ZAKKI

Zentrale Anlaufstelle für innovatives Lehren und Lernen interdisziplinärer Kompetenzen der KI

KI für alle

Künstliche Intelligenz Verstehen, Anwenden und Reflektieren

Prof. Dr. Sebastian von Enzberg

Magdeburg, 27.05.2024

h2.de/zakki

Hochschule Magdeburg • Stendal

GEFÖRDERT VOM Bundesministerium für Bildung und Forschung

Überblick: Themen und Daten

Termin	Nr	Themen *(vorläufig)*
08.04.2024	**01**	Was ist KI?
15.04.2024	**02**	KI Anwendungen
22.04.2024	**03**	KI Algorithmen: Grundfunktionen und wissensbasierte Verfahren
29.04.2024	**04**	KI Algorithmen: Deskriptive Statistik
06.05.2024	**05**	KI Algorithmen: Neuronale Netze und Machine Learning Paradigmen
13.05.2024	**06**	KI Algorithmen: Machine Learning Herausforderungen
~~*20.05.2024*~~		*- Keine Vorlesung (Feiertag) -*
27.05.2024	**07**	Datenrepräsentation und Datenqualität
~~*03.06.2024*~~		*- Keine Vorlesung (Selbststudium) -*
10.06.2024	**08**	Regulierung, Datenschutz und Datensicherheit
17.06.2024	**09**	IT Systeme für Big Data und KI
24.06.2024	**10**	KI Implikationen
01.07.2024	**11**	Zukunftsszenarien: Abschätzung künftiger Entwicklungen
08.07.2024	**12**	Zukunftsszenarien: Futurologie und Post-AI Humanism

Lernziele für heute

…die digitale Repräsentation von Daten für verschiedene Modalitäten **kennen** und deren Eigenschaften **verstehen**.

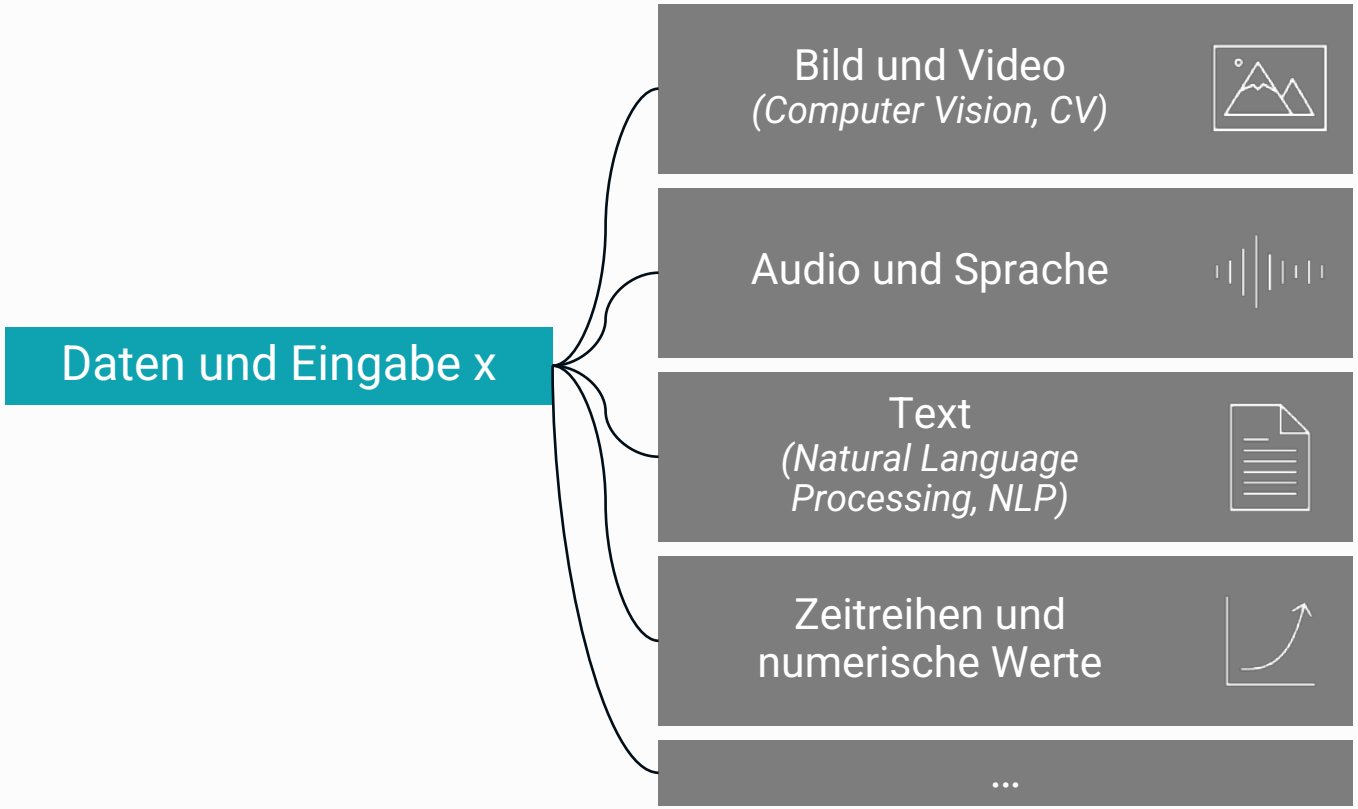
…die Bedeutung von manuell und automatisiert definierten Merkmalsräumen **kennen**, eigene Merkmale **definieren.**

…Datenquellen hinsichtlich Qualität, Verfügbarkeit und Kosten **bewerten**.

Datenrepräsentation und Datenqualität

1. **Datenrepräsentation**
2. Merkmale und Merkmalsräume
3. Datenquellen und Datenqualität

Wiederholung: Datenmodalitäten

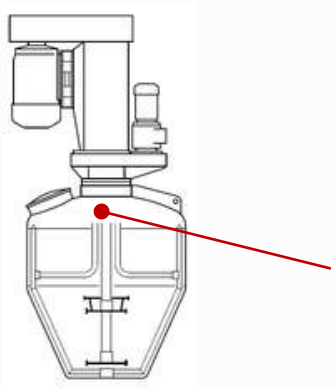


Datenmodalitäten

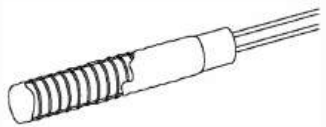
- **Strukturierte Daten** folgen einem standardisierten Format. Sie sind in Tabellen abbildbar und folgen definierten Datentypen (nominal, ordinal, numerisch), z.B. Messreihen.
- **Unstrukturierte Daten** folgen keinem festen Datenmodell, das semantische Information trägt, z.B. Bilder, Audio, Video, Texte. (Ein Datenformat beschreibt lediglich die technische Ablage.)
- **Meta-Daten** ergänzen oft Datensätze um beschreibende Eigenschaften (z.B. Datum der Speicherung oder Veränderung, Autor, Aufnahmeort).

- Viele Daten basieren auf **Signalen**, d.h. Darstellung von **messbaren physikalischen Größen** (z.B. Audio, Video, Bilder, Messreihen)

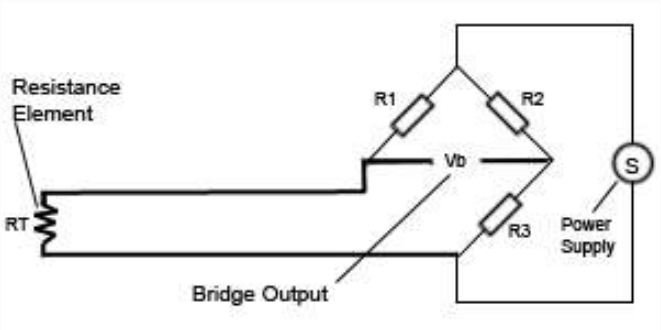
Messwertaufnahme



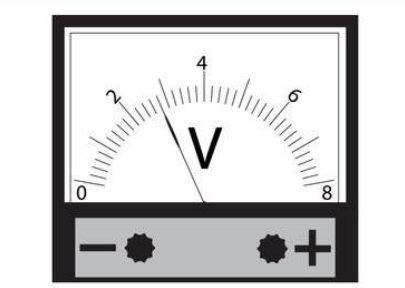
Maschine



Widerstands-thermometer

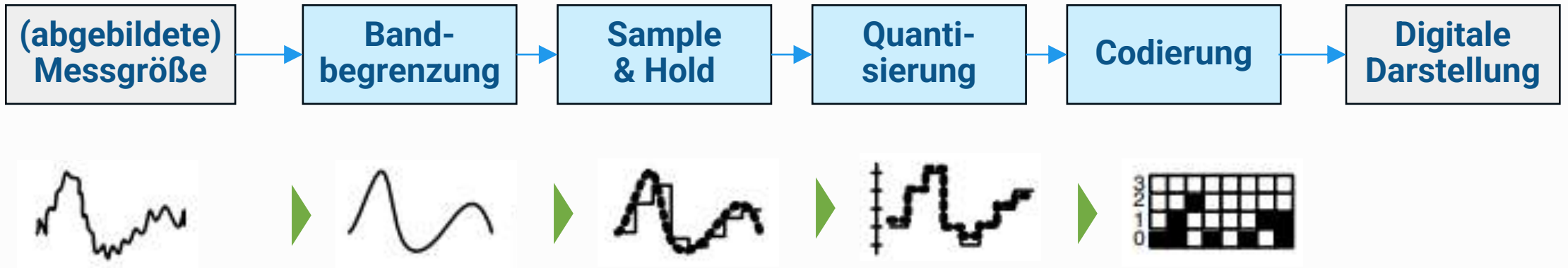


Brückenschaltung



Voltmeter

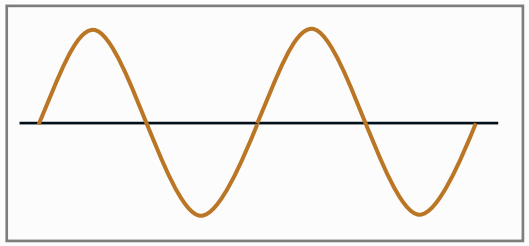
Digitalisierung



- Die Digitalisierung von Daten ist **Grundlage für Speicherung, Übertragung und automatisierte Verarbeitung** mit Computern.

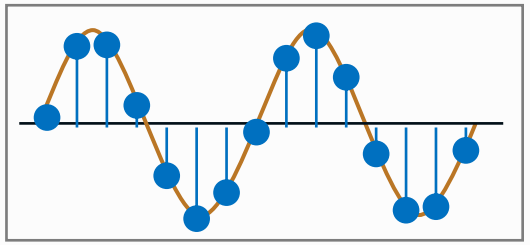
Digitalisierung: Zeitquantisierung

Signal (Beispiel)



Messung zu diskreten Zeitpunkten

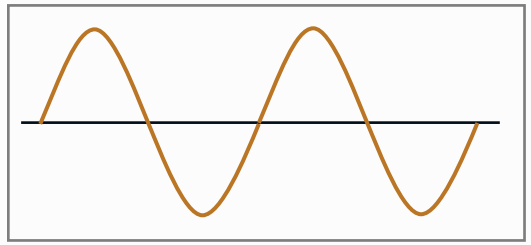
Abgetastetes Signal

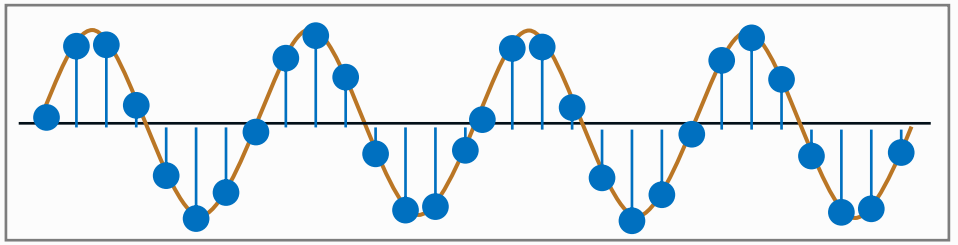


- Die **Abtastrate** beschreibt die Frequenz, mit der ein Signal regelmäßig erfasst wird (Anzahl pro Sekunde = Hertz [Hz])
- Um Informationsverlust zu vermeiden, muss die **Abtastrate mindestens das Doppelte der höchsten Frequenz** des Signals betragen.

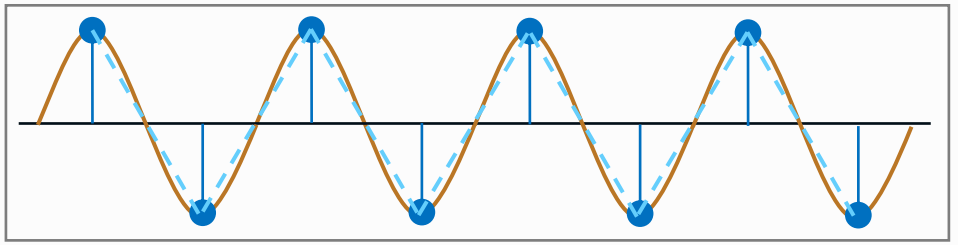
Digitalisierung: Zeitquantisierung

Signal (Beispiel)

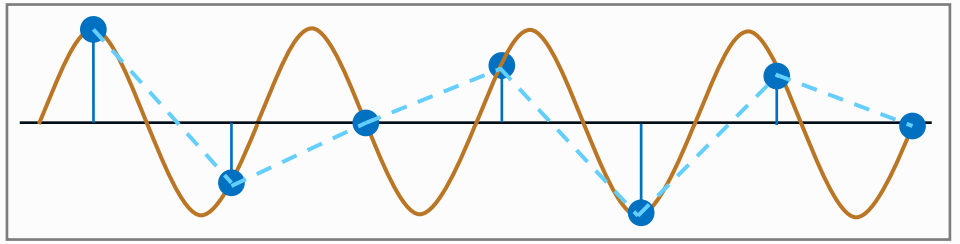




Hohe Abtastrate



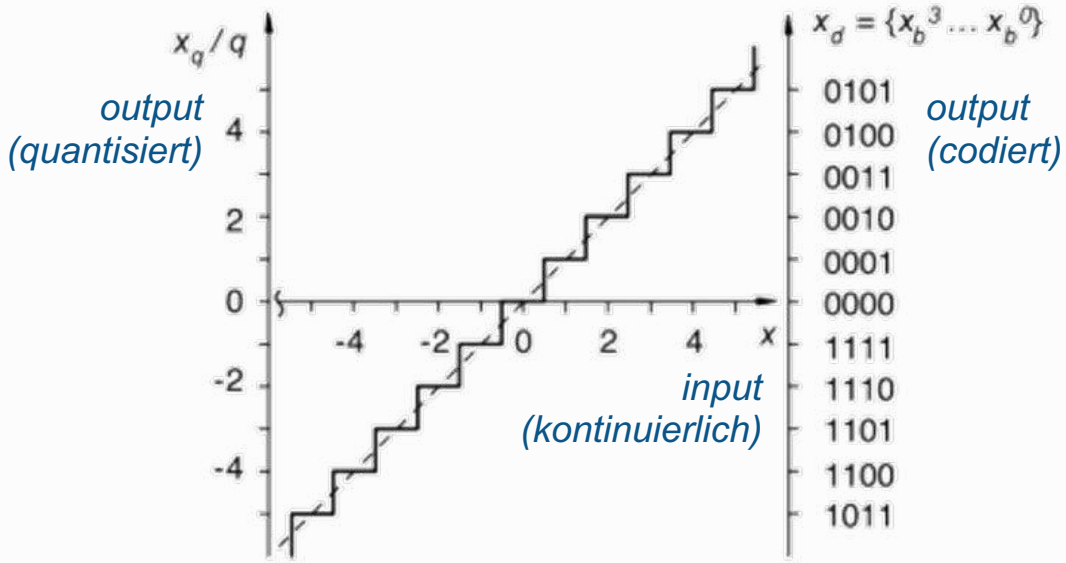
Minimale Abtastrate (2x Signalfrequenz)



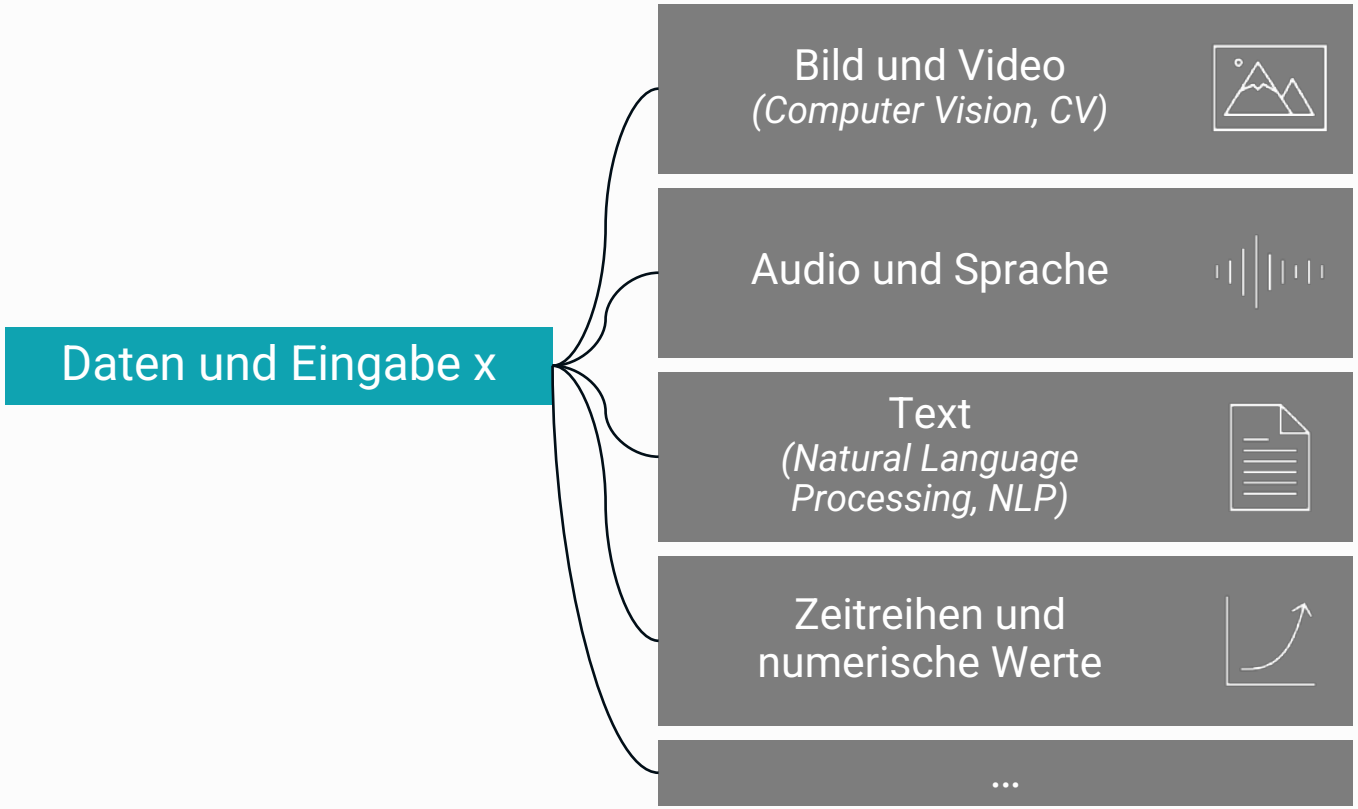
Unterabtastung

Digitalisierung: Amplitudenquantisierung

- Nach zeitlicher Abtastung müssen die Amplitudenwerte in eine **endliche Anzahl diskreter Werte** umgewandelt werden.
- Die **Bit-Tiefe** beschreibt die maximale Anzahl der Werte, die zur Darstellung des Signals verwendet werden können.
- (z.B. 8-Bit = 256 Werte; 16-Bit = 65 536 Werte)
- Eine **höhere Bit-Tiefe** liefert eine **feinere Auflösung** der Amplitudenwerte.
- Der verbleibende Fehler wird als **Quantisierungsfehler** bezeichnet.

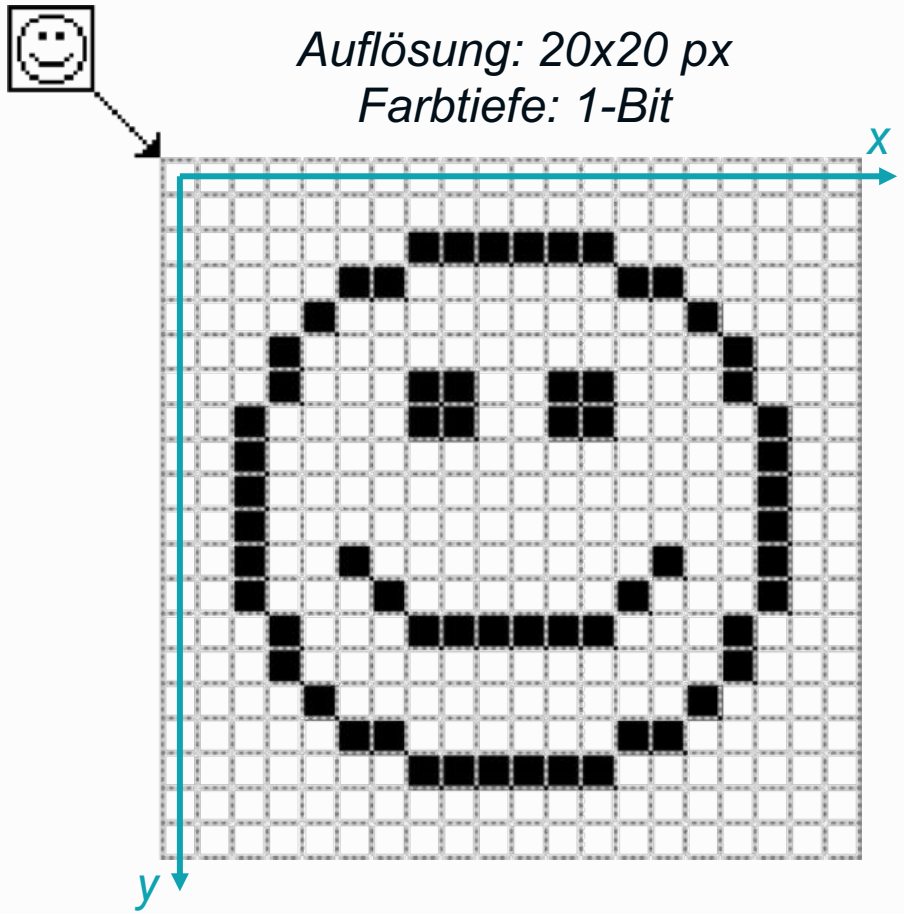


Wiederholung: Datenmodalitäten



Datenmodalität: Bilder

- Digitale Darstellung visueller Informationen als Helligkeitswerte in **2D-Matrix Anordnung**
- Ein **Pixel** ist dabei ein Helligkeitswert an einer Position x/y und die kleinste Einheit
- Die **Farbtiefe** bezeichnet die Bit-Tiefe eines Helligkeitswertes (z.B. Bild rechts ist ein Binärbild schwarz/weiß)



Datenmodalität: Bilder

- Mit 8-Bit können 256 Helligkeitswerte je Pixel dargestellt werden (von 0 = Schwarz bis 255 = weiß)
- Farbwerte ergeben sich als Kombination von Helligkeitswerten von Grundfarben (z.B. RGB = Rot + Grün + Blau mit je 8-Bit)

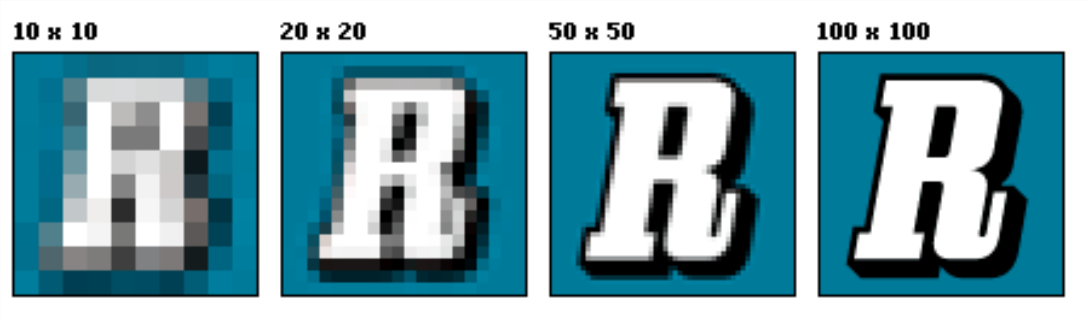
1-Bit

8-Bit Grautöne

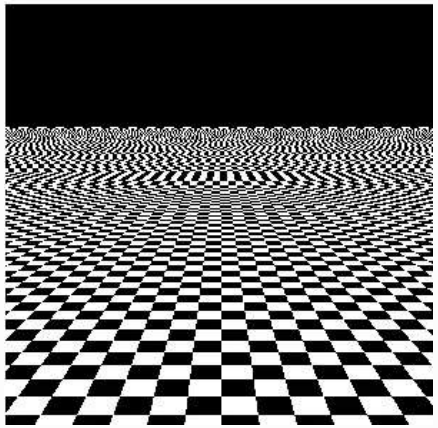
3x8-Bit Farbe

Datenmodalität: Bilder

- Die **Auflösung** bezeichnet die Anzahl der Pixel in Breite x Höhe (x, y)

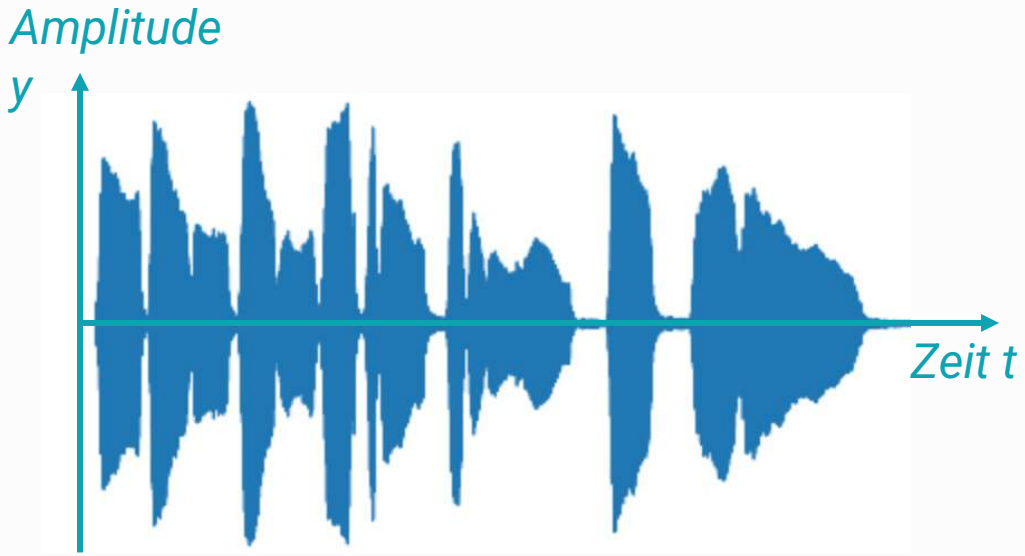


- Bei der örtlichen Abtastung gelte ähnliche Überlegungen wie bei der zeitlichen Abtastung:
- Die Auflösung muss gegenüber der Periodenlänge eines Musters hoch genug sein!



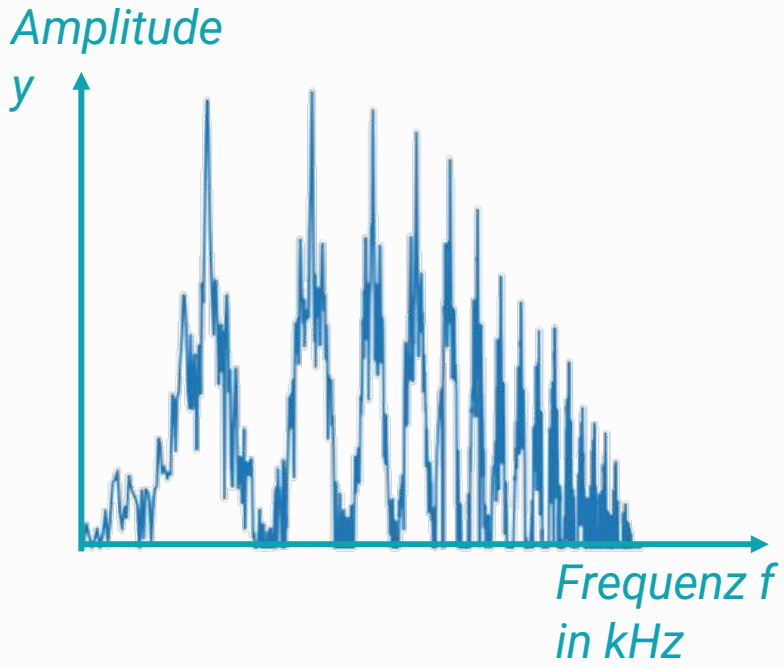
Datenmodalität: Audio

- Digitale Darstellung akustischer Information als **Zeitverlauf einer Amplitude**
- Typische **Abtastraten** sind z.B.
  44,1 kHz für CD-Qualität oder
  8 kHz einfache Sprachqualität (Telefon)
- Typische **Bit-Tiefe** liegt z.B. bei
  16 Bit (CD-Qualität)



Datenmodalität: Audio

- Das **Spektrum** eines Audio-Signals zeigt die Anteile von tiefen Tönen (niedrige Frequenz) bis zu hohen Tönen (hohe Frequenz)
- **Spektrale Merkmale** sind daher für die Audio-Verarbeitung besonders gut geeignet (*siehe Abschnitt Merkmale*)



Datenmodalität: Text

- Textdaten werden als **Sequenz von Zeichen** dargestellt

```
Natural Language
Processing (NLP)
```

```
4E 61 74 75 72 61 6C 20 4C 61 6E 67 75 61 67 65
50 72 6F 63 65 73 73 69 6E 67 20 28 4E 4C 50 29
```

- Zeichen können z.B. in **ASCII (8 Bit)** oder in **Unicode (32 Bit)** codiert werden

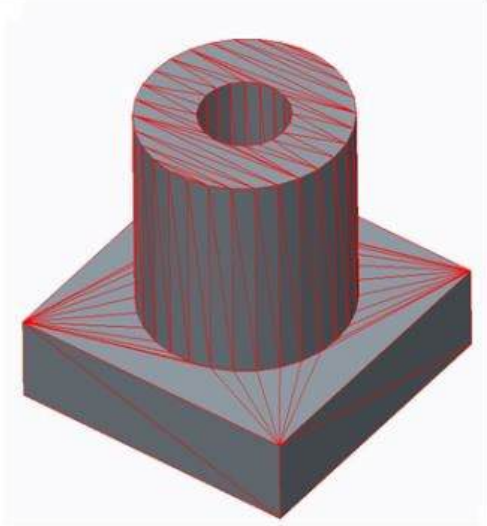
$b_7 b_6 b_5$ → / Bits	000	001	010	011	100	101	110	111
b_4 b_3 b_2 b_1 ↓ / Column → / Row ↓	0	1	2	3	4	5	6	7
0 0 0 0	0	NUL	DLE	SP	0	@	P	`	p
0 0 0 1	1	SOH	DC1	!	1	A	Q	a	q
0 0 1 0	2	STX	DC2	"	2	B	R	b	r
0 0 1 1	3	ETX	DC3	#	3	C	S	c	s
0 1 0 0	4	EOT	DC4	$	4	D	T	d	t
0 1 0 1	5	ENQ	NAK	%	5	E	U	e	u
0 1 1 0	6	ACK	SYN	&	6	F	V	f	v
0 1 1 1	7	BEL	ETB	'	7	G	W	g	w
1 0 0 0	8	BS	CAN	(	8	H	X	h	x
1 0 0 1	9	HT	EM	)	9	I	Y	i	y
1 0 1 0	A	LF	SUB	*	:	J	Z	j	z
1 0 1 1	B	VT	ESC	+	;	K	[	k	{
1 1 0 0	C	FF	FS	,	<	L	\	l	|
1 1 0 1	D	CR	GS	-	=	M	]	m	}
1 1 1 0	E	SO	RS	.	>	N	^	n	~
1 1 1 1	F	SI	US	/	?	O	_	o	DEL

ASCII-Zeichencodierung

Datenmodalität: Text

- Textdaten sind Grundlage für viele **abgeleitete Datentypen** (semi-strukturierte und strukturierte), z.B.
  - Markdown, XML, JSON
  - Programmiersprachen
  - Logs und Skripte
  - STL, SVG, …

```
solid CREO STL
 facet normal 0.000000e+00 0.000000e+00 -1.000000e+00
  outer loop
   vertex 5.000000e+01 -5.000000e+01 0.000000e+00
   vertex -5.000000e+01 5.000000e+01 0.000000e+00
   vertex 5.000000e+01 5.000000e+01 0.000000e+00
  endloop
 endfacet
 facet normal 0.000000e+00 1.000000e+00 0.000000e+00
  outer loop
   vertex 5.000000e+01 5.000000e+01 6.000000e+00
   vertex 5.000000e+01 5.000000e+01 0.000000e+00
   vertex -5.000000e+01 5.000000e+01 0.000000e+00
  endloop
 endfacet
.....
.....
endsolid CREO STL
```



STL Beschreibung für 3D-Objekt

- Eine Codierung in größere **Zeichen-blöcke („Tokens“)** ist für die weitere Verarbeitung besonders gut geeignet (*siehe Abschnitt Merkmale*)

Text can be coded in tokens for better application in Natural Language Processing.

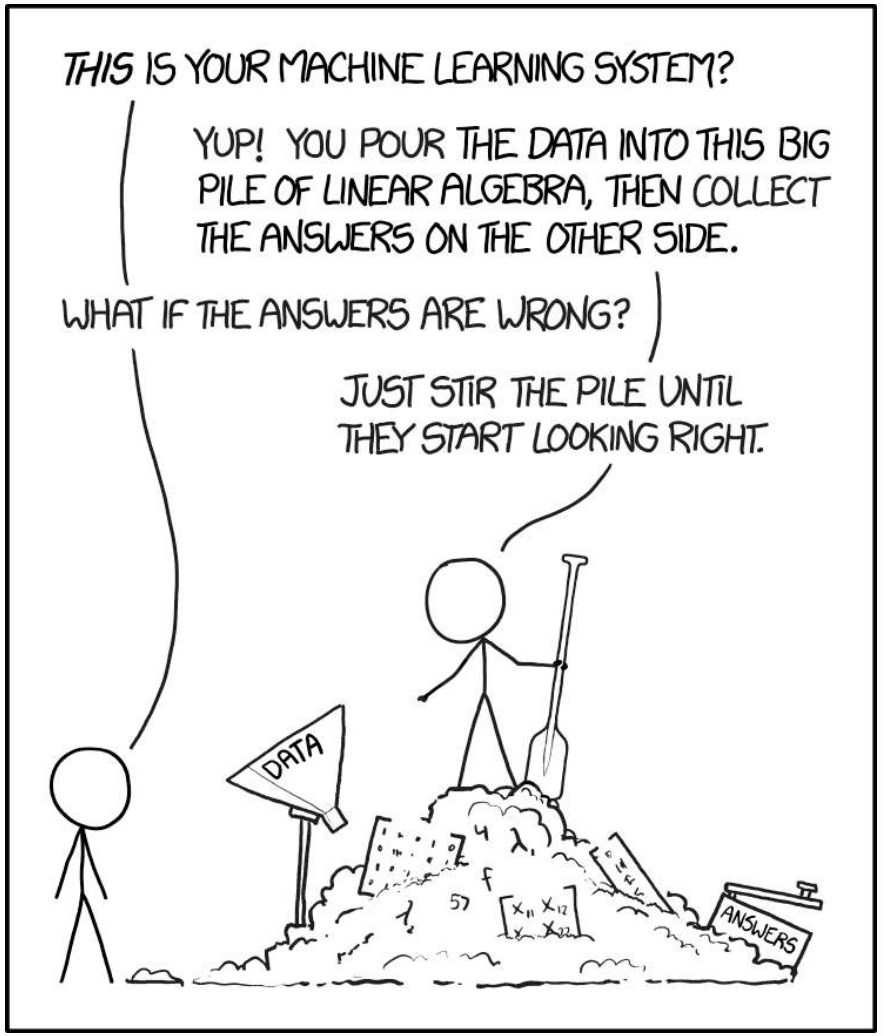
https://platform.openai.com/tokenizer

Datenrepräsentation und Datenqualität

1. Datenrepräsentation
2. **Merkmale und Merkmalsräume**
3. Datenquellen und Datenqualität

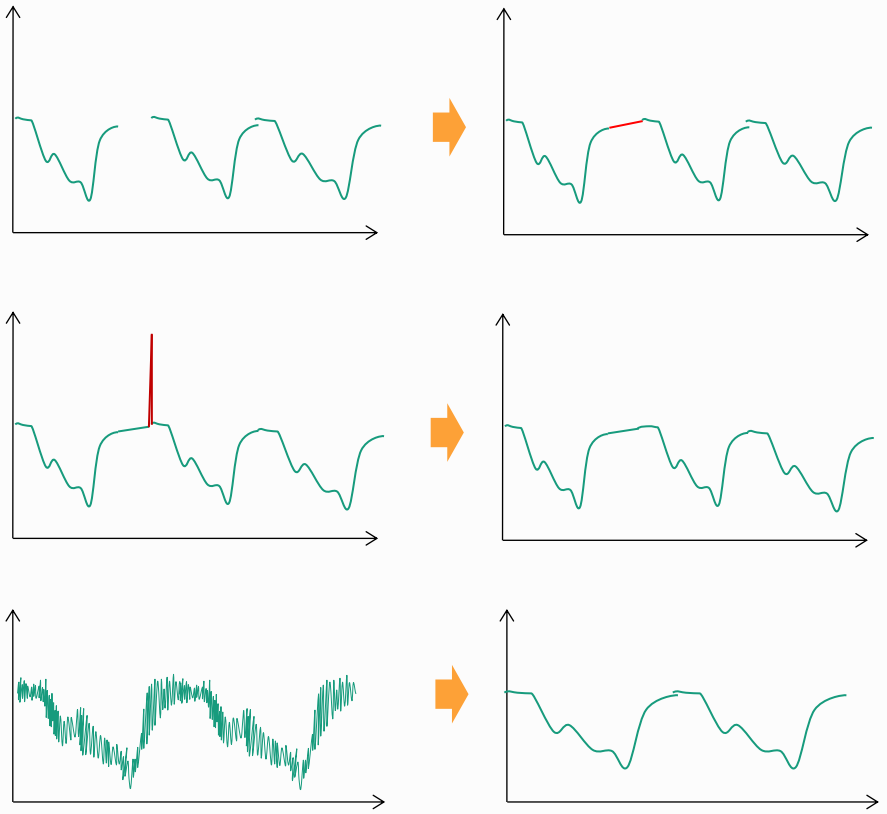
Merkmale

- **Merkmale *(Features)*** sind eine Form der Datenrepräsentation, die **besonders geeignet als Eingabe-Daten** für ein Modell sind
- Beispiele:
  - Einfache Vorverarbeitung der Daten
  - Auswahl einer Untermenge an Datensätzen oder Datenreihen (*Feature Selection / Reduction*)
  - Mathematische oder statistische Berechnung
  - Mathematische Transformationen, z.B. spektrale Transformation, Reihenentwicklung
  - Gelernte Repräsentationen (Representation Learning)



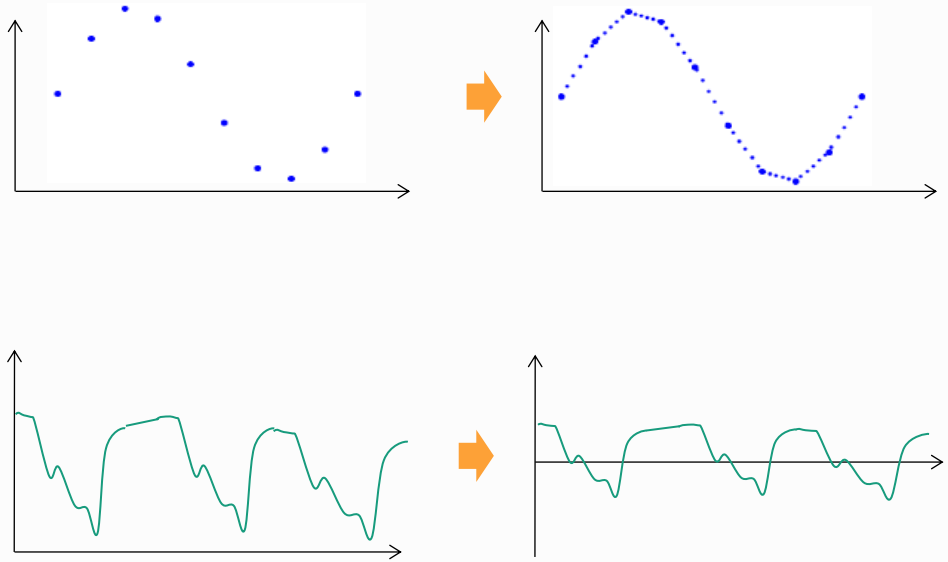
Datenvorverarbeitung

- Eine Aufgabe der Datenvorverarbeitung ist die Erhöhung der **Datenqualität**, z.B. durch
  - Behandlung Fehlender Werte
  - Behandlung von Ausreißern
  - Filtern von Signalen



Datenvorverarbeitung

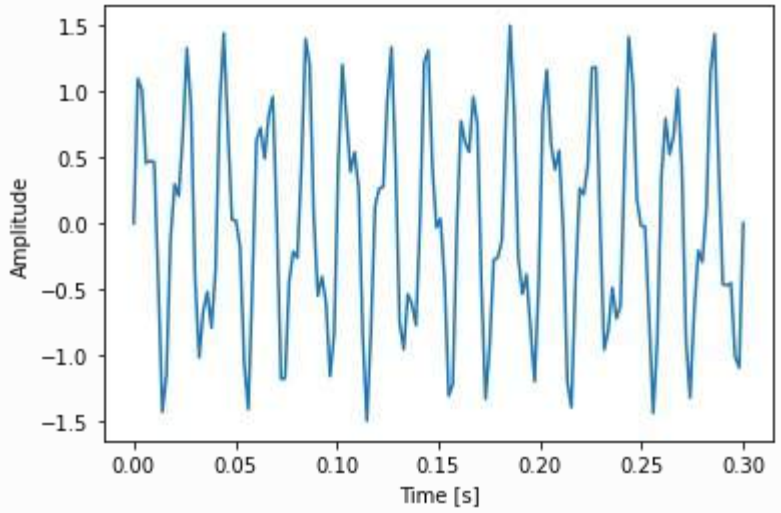
- Weiterhin kann durch Vorverarbeitung die Kompatibilität von Daten gewährleistet werden, bspw. durch
  - Resampling (z.B. bei Messungen mit verschiedener zeitl. Abtastung)
  - Normalisierung (z.B. bei stark unterschiedlichen Wertebereichen)



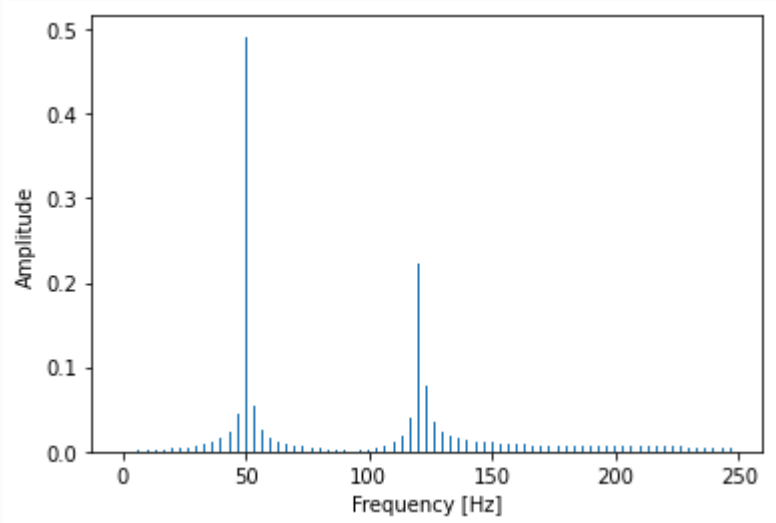
Spektrale Transformation

- Eine Spektrum ist das Ergebnis einer **Zerlegung eines Signals** in Bestandteile verschiedener Frequenzen
- Zerlegung basiert auf einer mathematischen **Faltungsoperation (Convolution)**

Zeitbereich



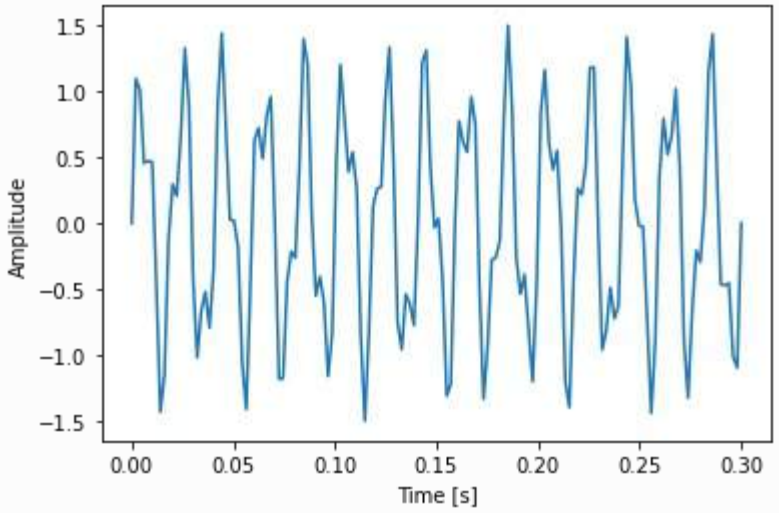
Frequenzbereich



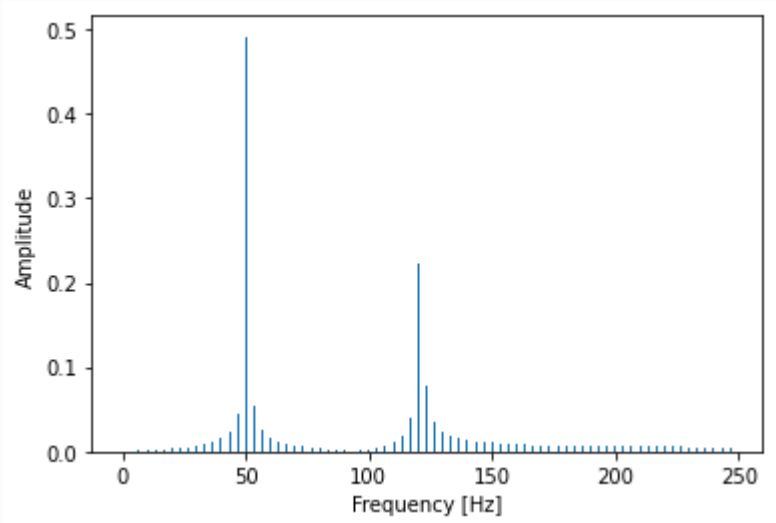
Spektrale Transformation

- Die **Fourier-Transformation** ist eine Zerlegung auf Sinus-/Kosinus-Bestandteile
- Ergebnis der Transformation hat oft einen höheren **Informationsgehalt** und ist daher besser geeignet für die Modellbildung

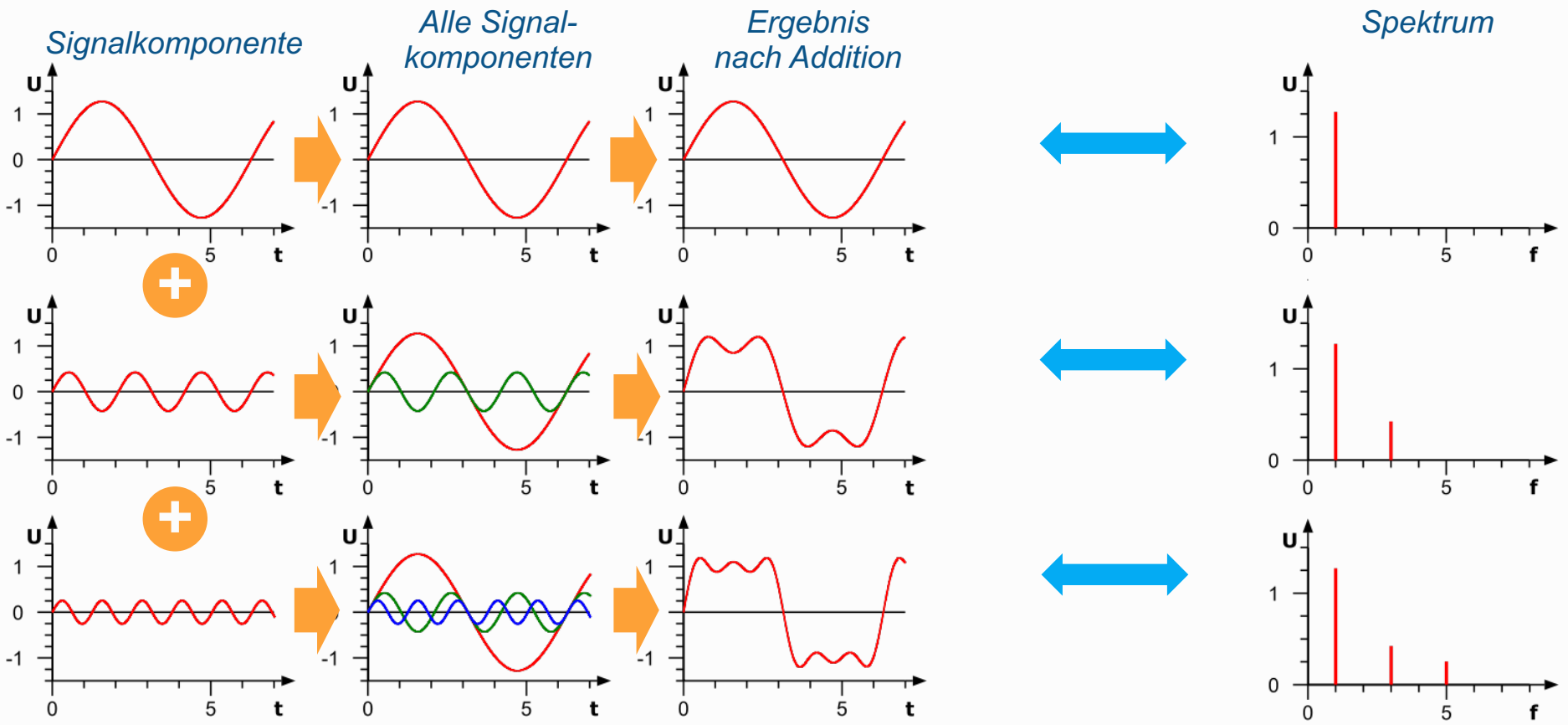
Zeitbereich (f_s = 500 Hz)



Frequenzbereich

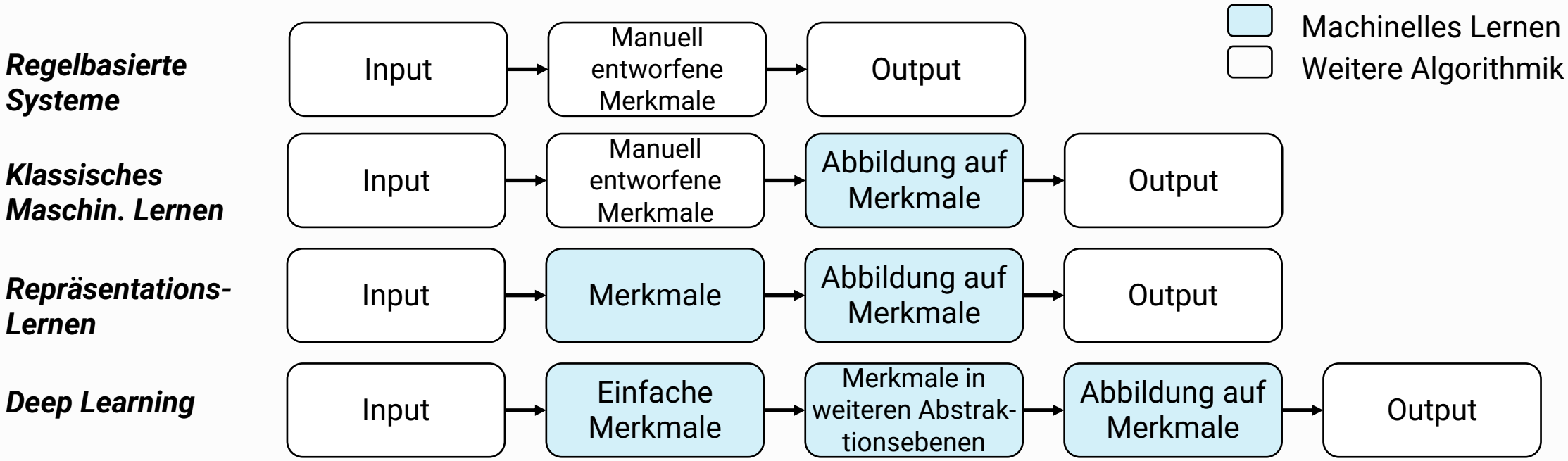


Spektrale Transformation: Beispiel



Representation Learning

- Representation Learning beschäftigt sich mit der **automatisierten Bestimmung geeigneter Merkmale** aus Daten

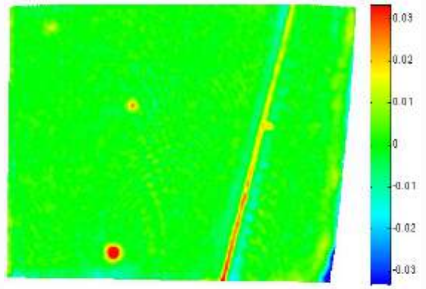


Bilder: Spezifische Merkmale

- In der Bildverarbeitung gibt es eine ganze Reihe von Operationen zur **Extraktion allgemeingültiger Bildmerkmale**
- **Flächenmerkmale**, z.B. Schwerpunkt-koordinaten, Flächeninhalt, Ausrichtung, Form
- **Kantenmerkmale**, z.B. Geraden-gleichungen, Winkel, Linienabschnitte
- **Punktmerkmale**, z.B. Koordinaten, Skalierung

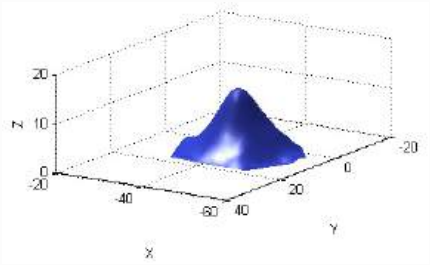
Bilder: Beispiel spezifische Merkmale

- Daneben können **anwendungsfalls-pezifische Merkmale** definiert werden
- **Beispiel:** Erkennung von Oberflächendefekten

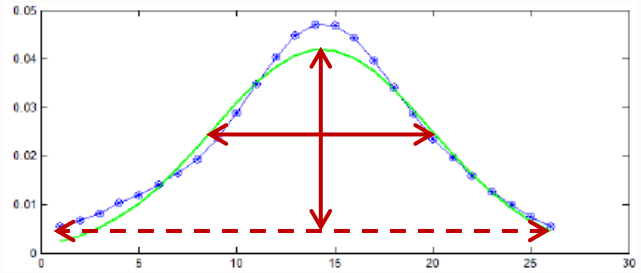


Defektkarte

2D-Ansicht Oberflächendefekt

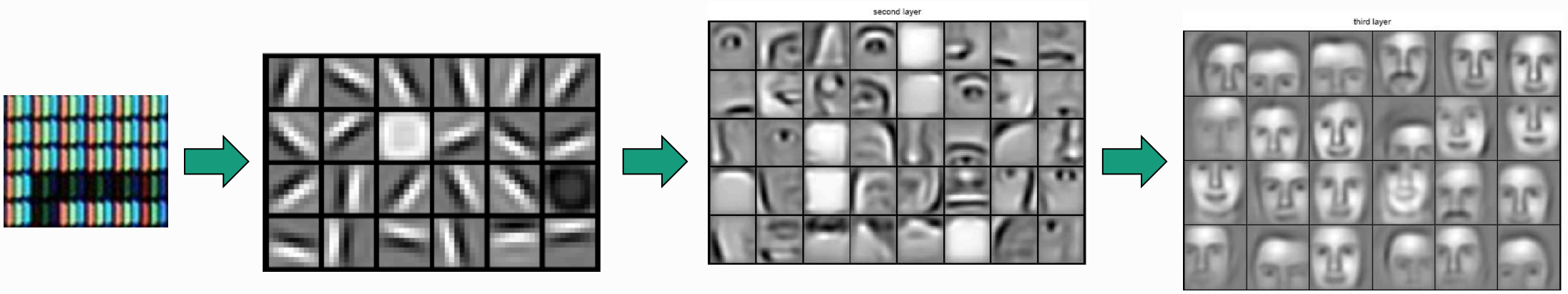


3D- Ansicht Oberflächendefekt



Bilder: Representation Learning

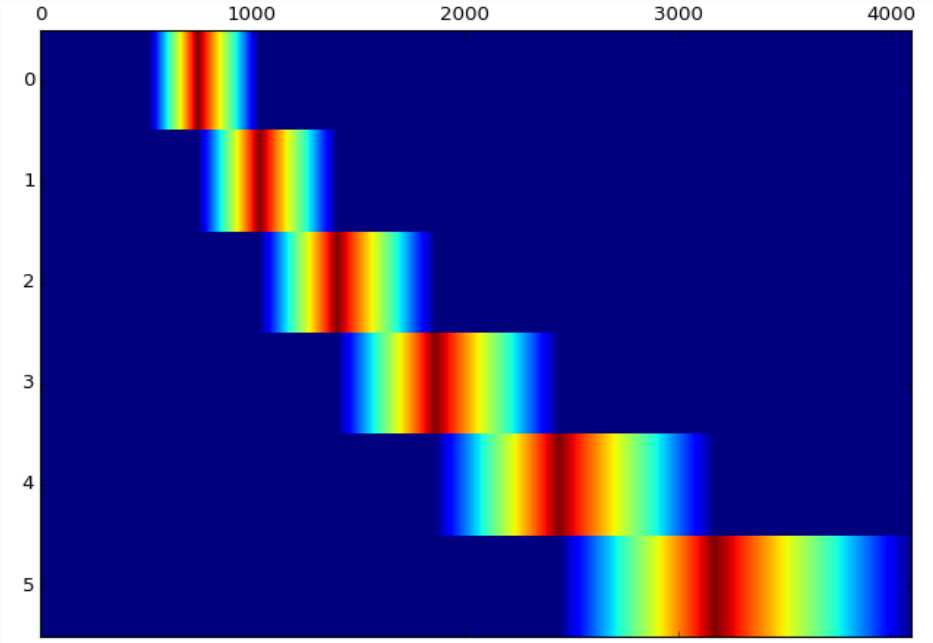
- Künstliche Neuronale Netze, insbesondere **Convolutional Neural Networks (CNNs)** sind sehr gut geeignet, automatisiert Merkmale aus Bilddaten zu bestimmen



Pixel | Edges | Object parts | Object models

Audio: Spezifische Merkmale

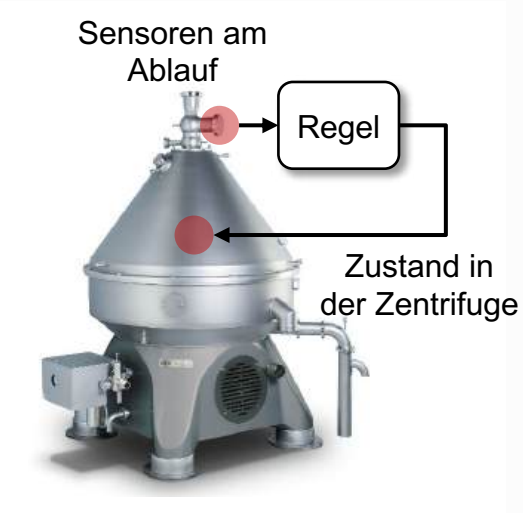
- **Zeitdomänenmerkmale**, z.B. Signalenergie, maximale Amplitude, Verteilungsmerkmale
- **Frequenzdomänenmerkmale**, z.B. Fourier-Koeffizienten, Mel-Frequenz Cepstrum Koeffizienten (MFCCs)
- **Zeit-Frequenz-Domäne**, z.B. Spektogramm
- **Musikklassifikation**, z.B. Tonart, Tempo oder Klangfarbe-Merkmale für Genre-Erkennung



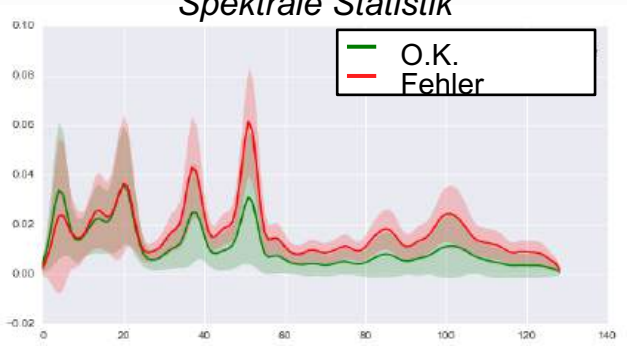
Mel Filterbank

Audio: Beispiel spezifische Merkmale

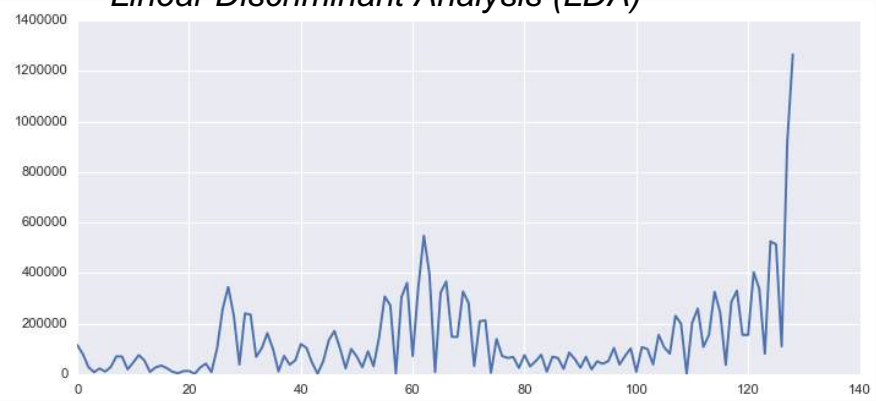
- **Beispiel:** Erkennung eines Fehlerzustandes einer Zentrifuge aufgrund von Vibrationen



Spektrale Statistik



Linear Discriminant Analysis (LDA)



Text: Spezifische Merkmale

- Merkmale in der Textverarbeitung (NLP) basieren oft auf statistischen Häufigkeiten von Worten
- **N-Grams:** Wahrscheinlichkeit für Aufeinanderfolge von 2, 3,… Wörtern
- **TF-IDF:** Term Frequency-Inverse Document Frequency (Häufigkeit in einem Dokument relativ zu einem Testkorpus)
- **Statistische Beschreibungen:** Wortanzahl, Satzanzahl, durchschnittliche Satzlänge

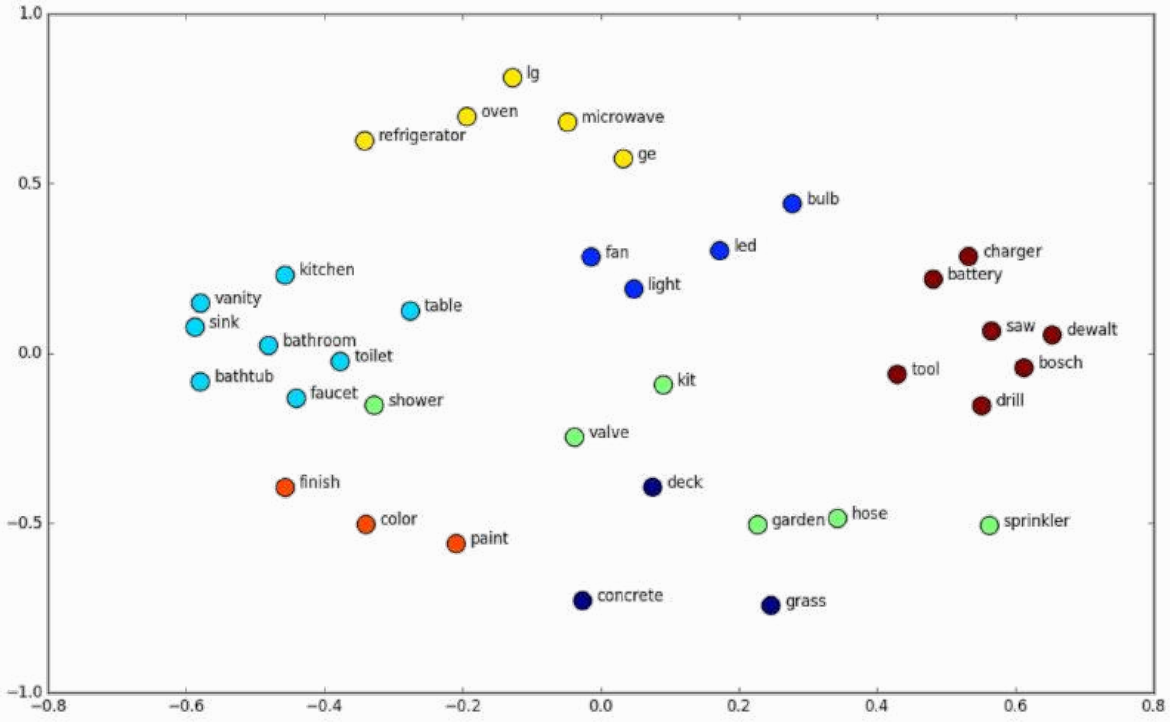
Text: Representation Learning

- **Word Embeddings** sind vektorielle Darstellungen einzelner Worte oder Tokens, die durch Neuronale Netze gelernt wurden.

Text can be coded in tokens for better application in Natural Language Processing.

https://platform.openai.com/tokenizer

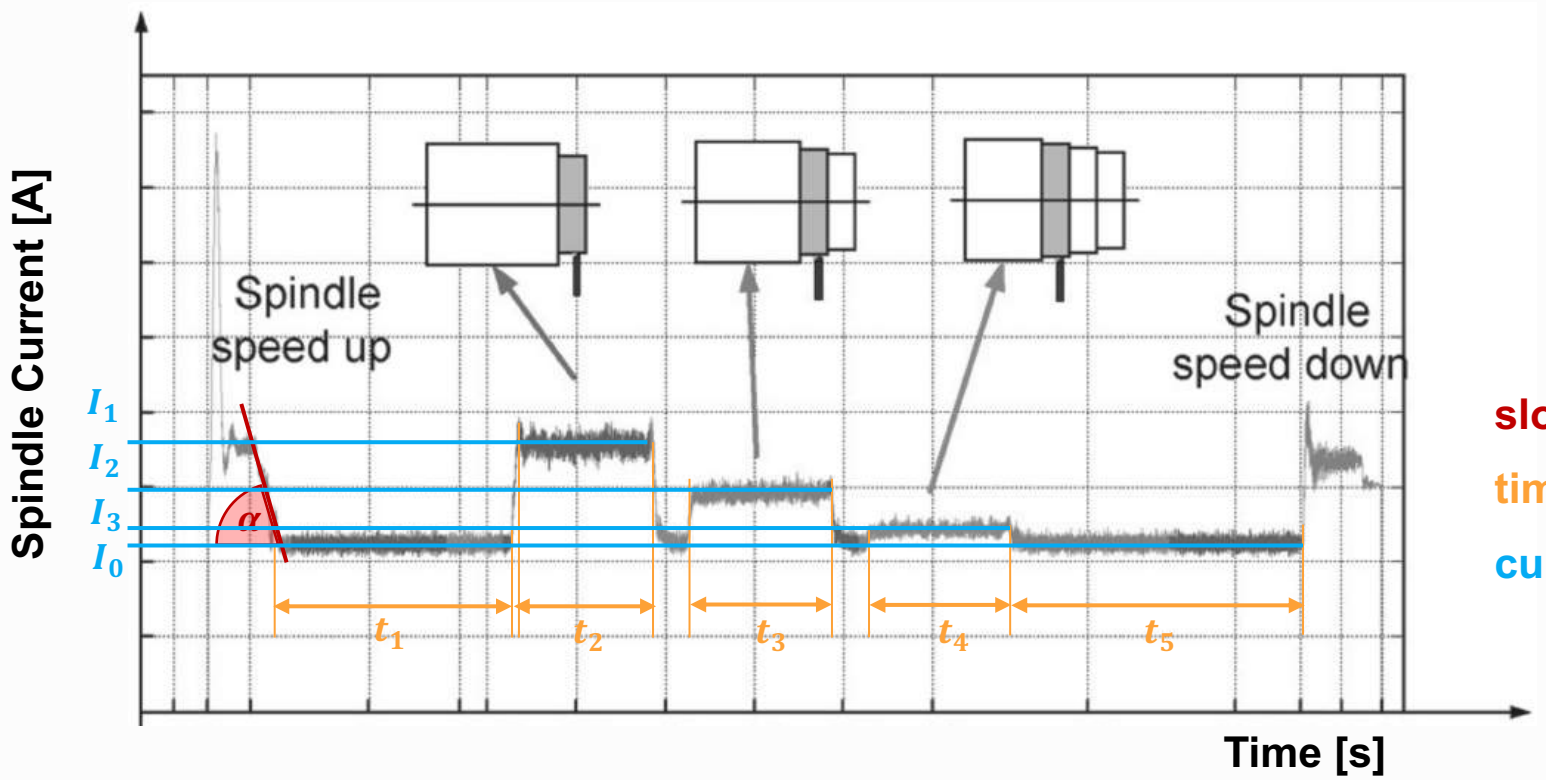
[1199, 649, 387, 47773, 304, 11460, 369, 2731, 3851, 304, 18955, 11688, 29225, 13]



https://projector.tensorflow.org

Zeitreihen: Beispiel spezifische Merkmale

- **Beispiel:** Prozessbeschreibung an einer Drehmaschine



slope / angle

time durations

current plateau values

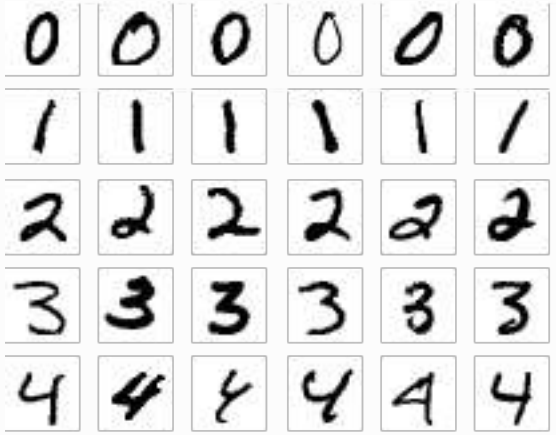
Datenrepräsentation und Datenqualität

1. Datenrepräsentation
2. Merkmale und Merkmalsräume
3. **Datenquellen und Datenqualität**

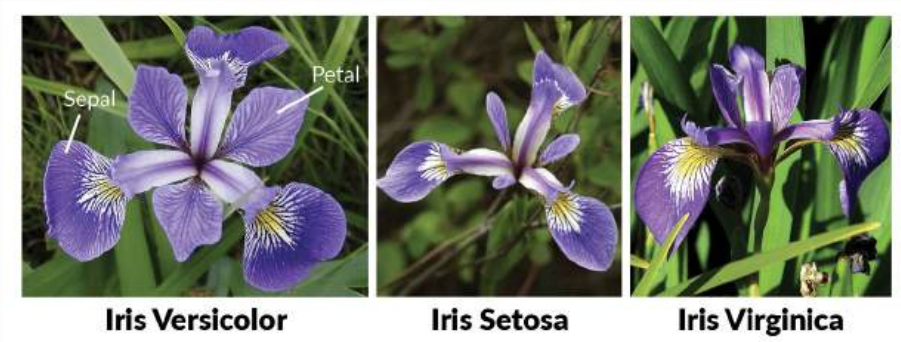
Datensätze für das Training

- Für viele Teilprobleme des maschinellen Lernens gibt es **frei verfügbare Datensätze**

MNIST: Erkennung handgeschriebener Zahlen



Iris: Erkennung von Blumen-sorten aufgrund von Merkmalen



- **Weitere Datensätze:** https://www.kaggle.com/datasets

Exkurs: Datenquellen in Unternehmen

ERP
Enterprise Resource Planning

PPC
Production Planning and Control

MES
Manufacturing Execution System

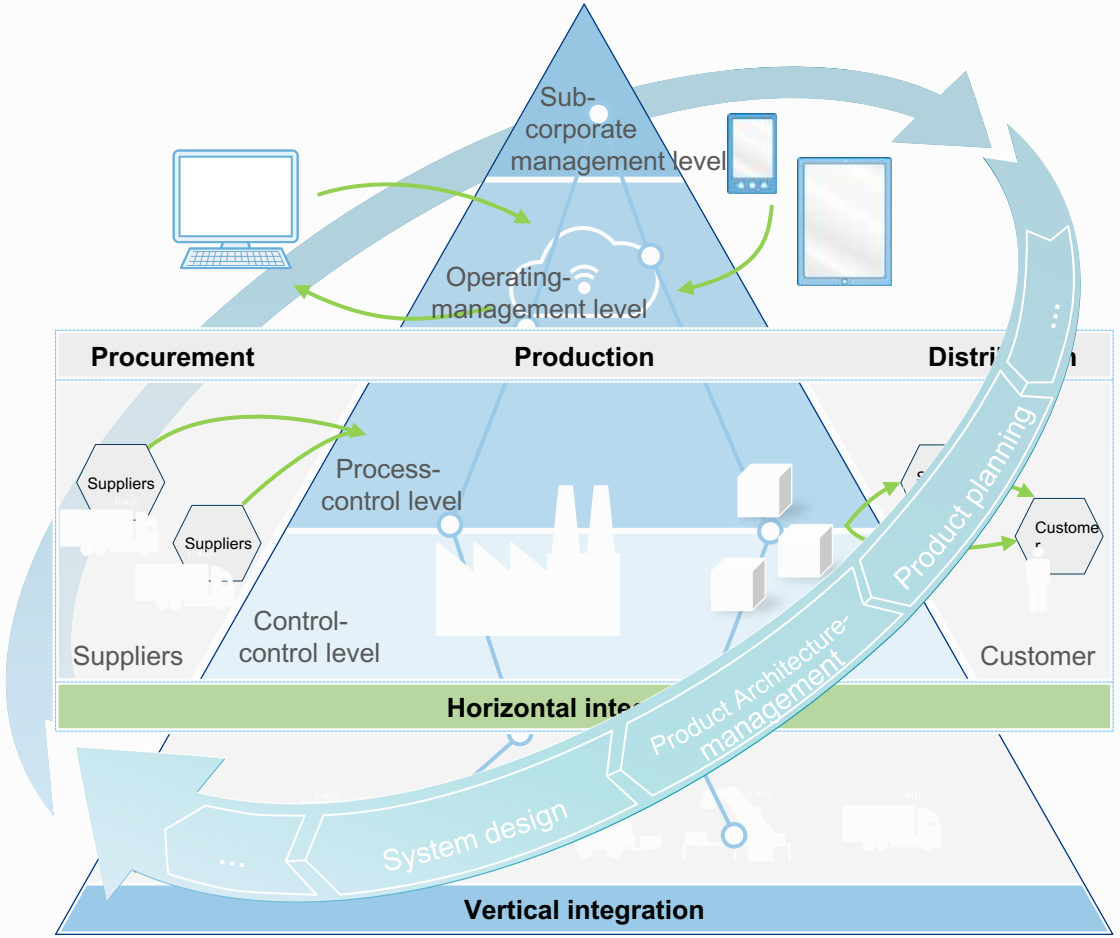
CAQ
Computer Aided Quality

CMMS
Comp. Maintenance Mgmt. Sys.

SCADA
Supervisory Ctrl. & Data Acquisition

DNC
Direct Numerical Control

PLC
Programmable Logic Controllers



CAD
Computer Aided Design

DMS
Document Management System

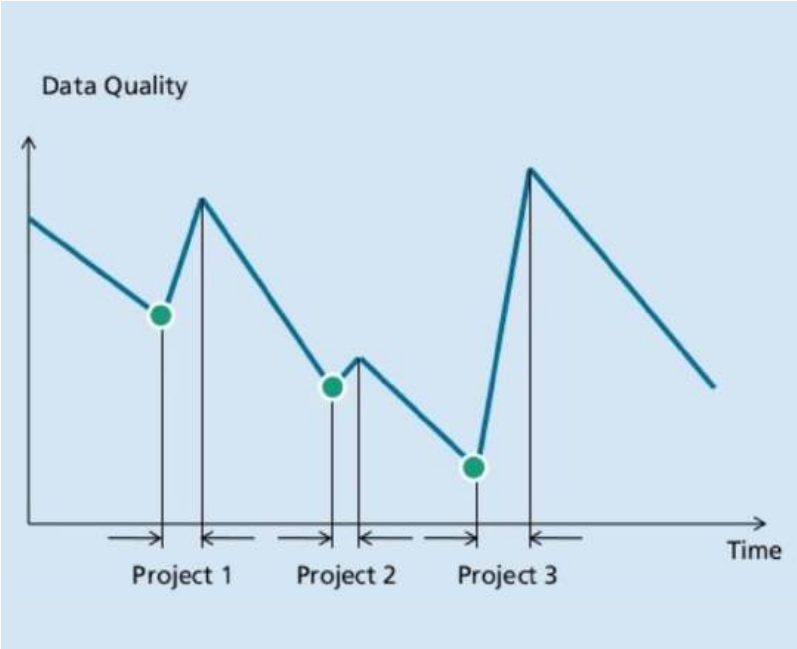
PLM
Product Lifecycle Management

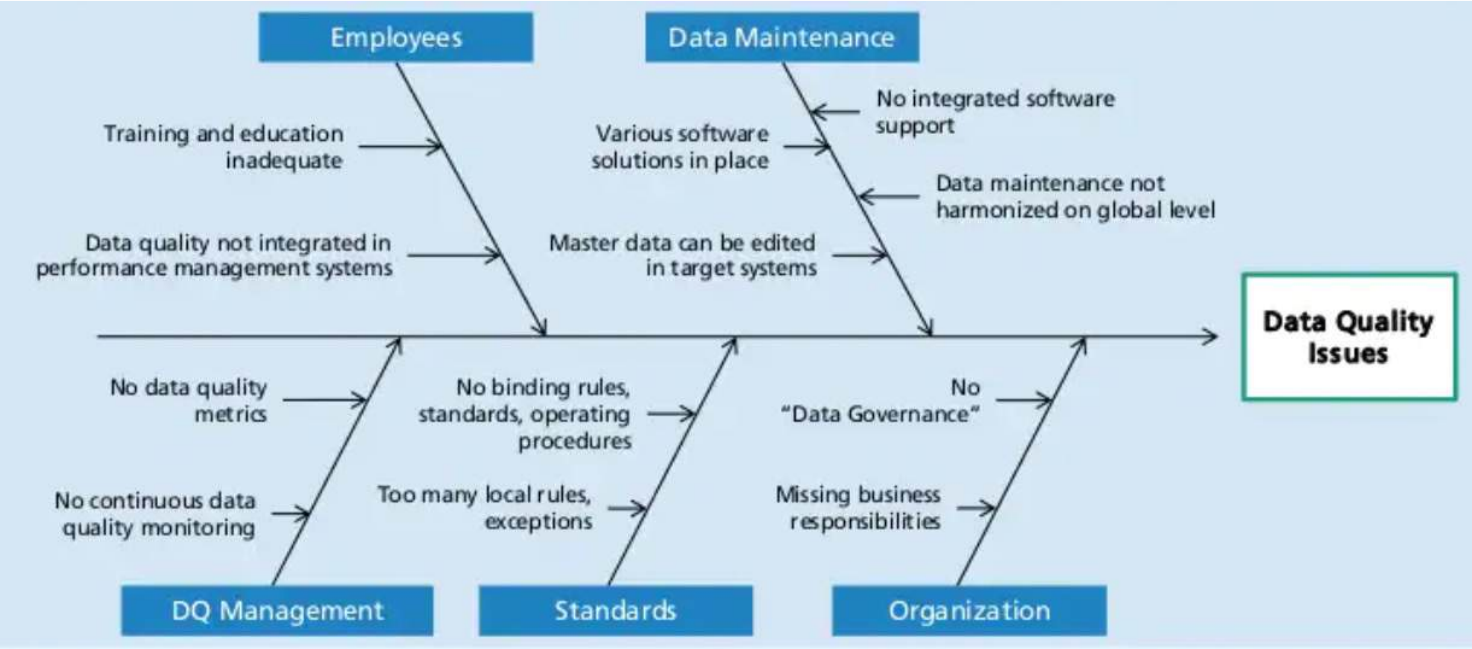
CRM
Customer Relationship Management

SCM
Supply Chain Mgmt.

BI
Business Intelligence

Datenqualität



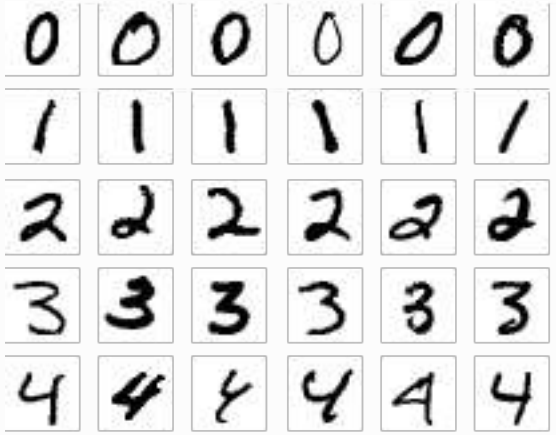


Quelle: Otto and Österle: Corporate Data Quality, 2016

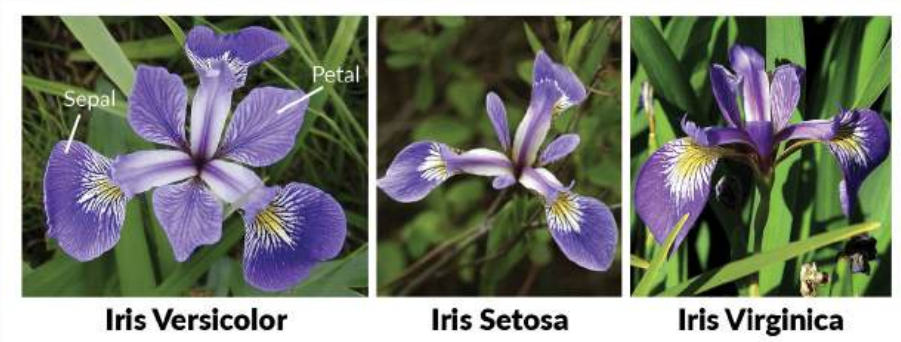
Datensätze für das Training

- Für viele Teilprobleme des maschinellen Lernens gibt es **frei verfügbare Datensätze**

MNIST: Erkennung handgeschriebener Zahlen



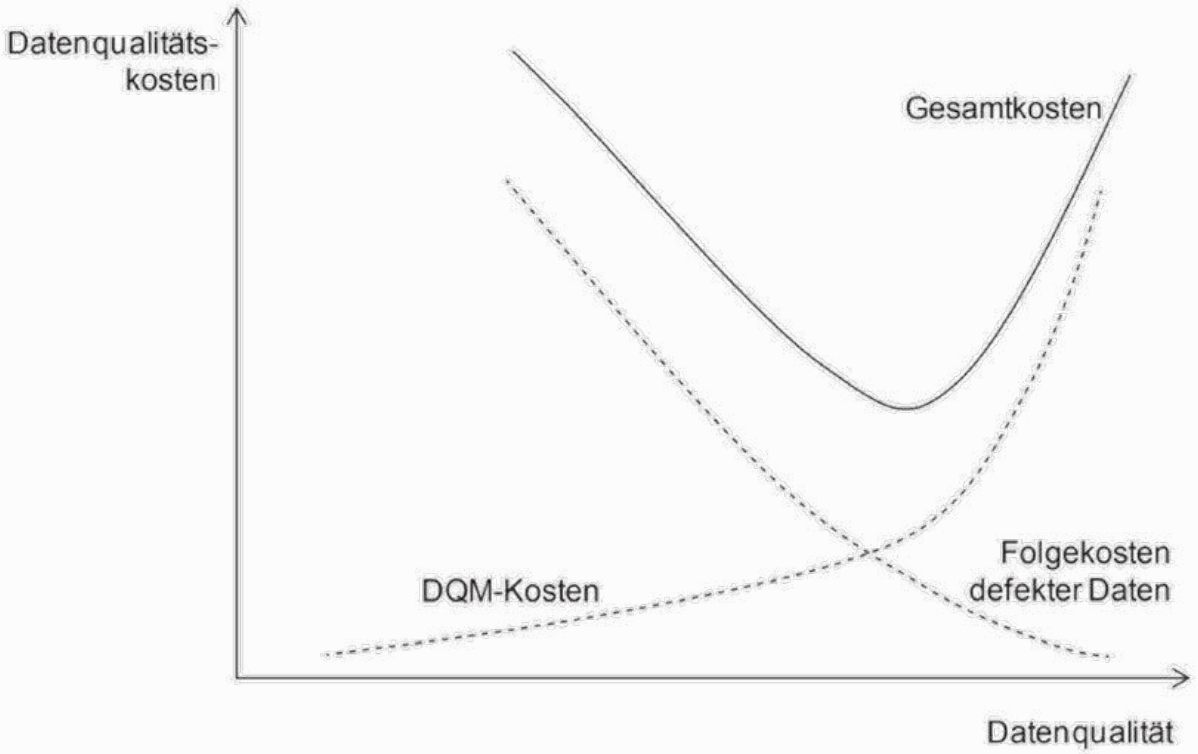
Iris: Erkennung von Blumen-sorten aufgrund von Merkmalen



- **Weitere Datensätze:** https://www.kaggle.com/datasets

Datenqualität

- Aufgabe eines Daten-Qualitäts-Managements ist die Gewährleistung einer **kontinuierlich hohen Datenqualität**
- Insb. wichtig bei **Modelldrift** und **Modellupdates**



Quelle: Otto and Österle: Corporate Data Quality, 2016

Lernziele für heute

…die digitale Repräsentation von Daten für verschiedene Modalitäten **kennen** und deren Eigenschaften **verstehen**.

…die Bedeutung von manuell und automatisiert definierten Merkmalsräumen **kennen**, eigene Merkmale **definieren.**

…Datenquellen hinsichtlich Qualität, Verfügbarkeit und Kosten **bewerten**.

Übung: Datenrepräsentation und Datenqualität

Dokumentieren Sie die Aufgaben im Lerntagebuch!

1. Erklären Sie anhand des Themas Ihrer Gruppe, **welche automatisiert definierten Merkmale** (Representation Learning) und **welche manuell definierten Merkmale** sie nutzen können, um ihre Modellqualität zu gewährleisten und zu verbessern!
2. Welche **Datenquellen für das Modelltraining** können Sie für Ihr Thema nutzen? Recherchieren Sie echte, verfügbare Datensätze und fassen Sie die wichtigsten Eigenschaften zusammen (5 bis max. 10 wichtigste Eigenschaften).
3. Welche **Datenquellen** würden Sie zukünftig aufnehmen, **um die Erkennungsleistung oder die Funktionalität zu verbessern**? Geben Sie eine grobe Abschätzung für den Aufwand zur Aufnahme der Daten, Herausforderungen und den erwarteten Nutzen!