

Data und Text Mining für systematische Forschungssynthesen

Alexander Christ

1. Data und Text Mining
2. Anwendung von Text Mining in Reviewverfahren am Beispiel der Digitalisierung in der kulturellen Bildung
3. Ausblick und Diskussion

- Data Mining und Text Mining sind Sammelbegriffe für Verfahren, die es ermöglichen mit „**Big Data**“ umzugehen.
- „**Big Data**“ bezeichnet Daten mit hoher Volume, Variety und Velocity. Dazu zählen auch Ergebnisse von Literaturrecherchen.
- Das Ziel der Verfahren ist, aus großen und ungeordneten Mengen an Daten Erkenntnisse zu gewinnen; bei Forschungssynthesen beispielsweise, welche Arbeiten eine höhere Wahrscheinlichkeit haben, inkludiert oder exkludiert zu werden.
- Die Anwendung von Text Mining bei Forschungssynthesen hat eine bereits „längere“ Geschichte (siehe O'Mara-Eves et al. (2015). Using Text Mining for study identification in systematic reviews: a systematic review of current approaches. Systematic Reviews, 4, 1-22. <https://doi.org/10.1186/2046-4053-4-5>)

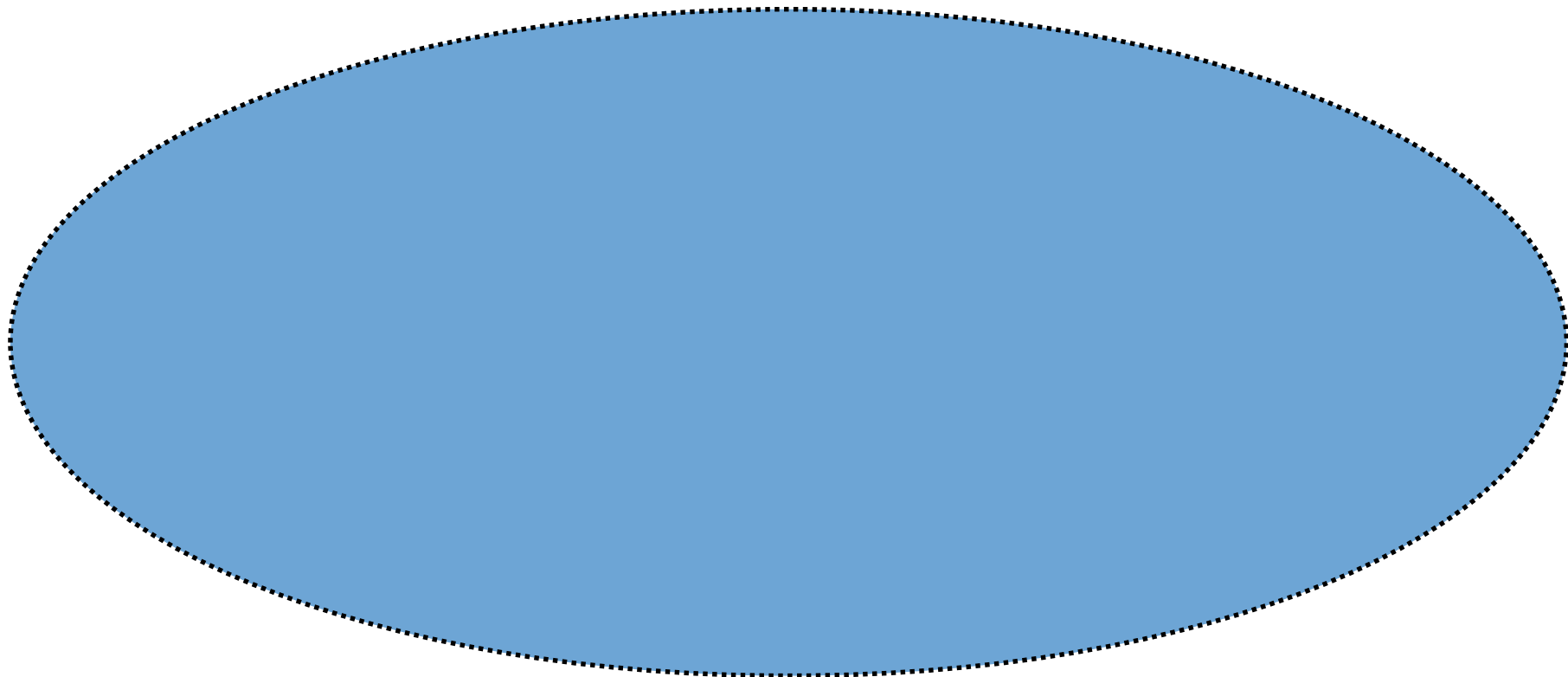
Anwendungsbeispiele für Text Mining bei Forschungssynthesen für die zentralen Schritte:

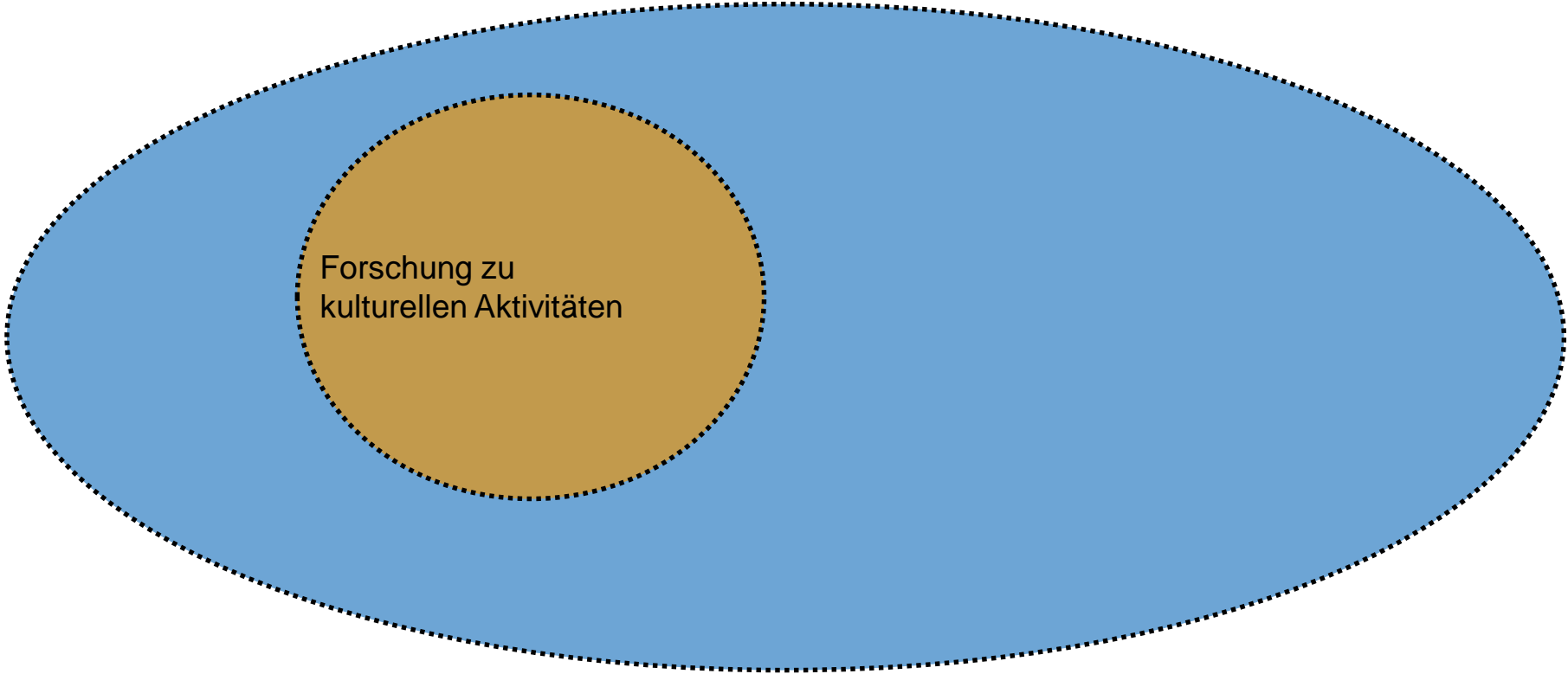
1. Suchbefehl und Literaturrecherche
2. Aufbereitung Korpus
3. Identifikation relevanter Arbeiten
4. Analyse und Kategorisierung

➤ **Beispiel der Digitalisierung in der kulturellen Bildung (DiKuBi)**
(siehe Christ et al., 2021, 2024)

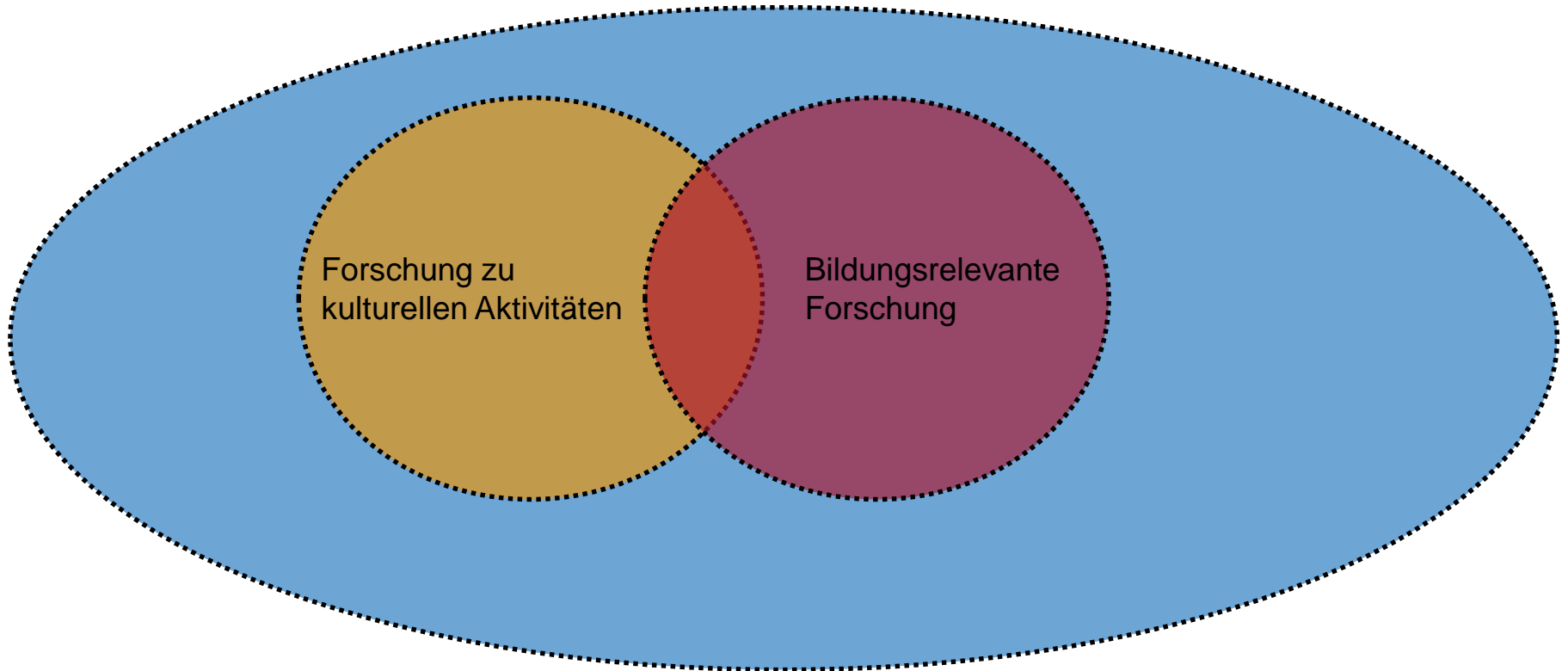
Der Forschungsgegenstand

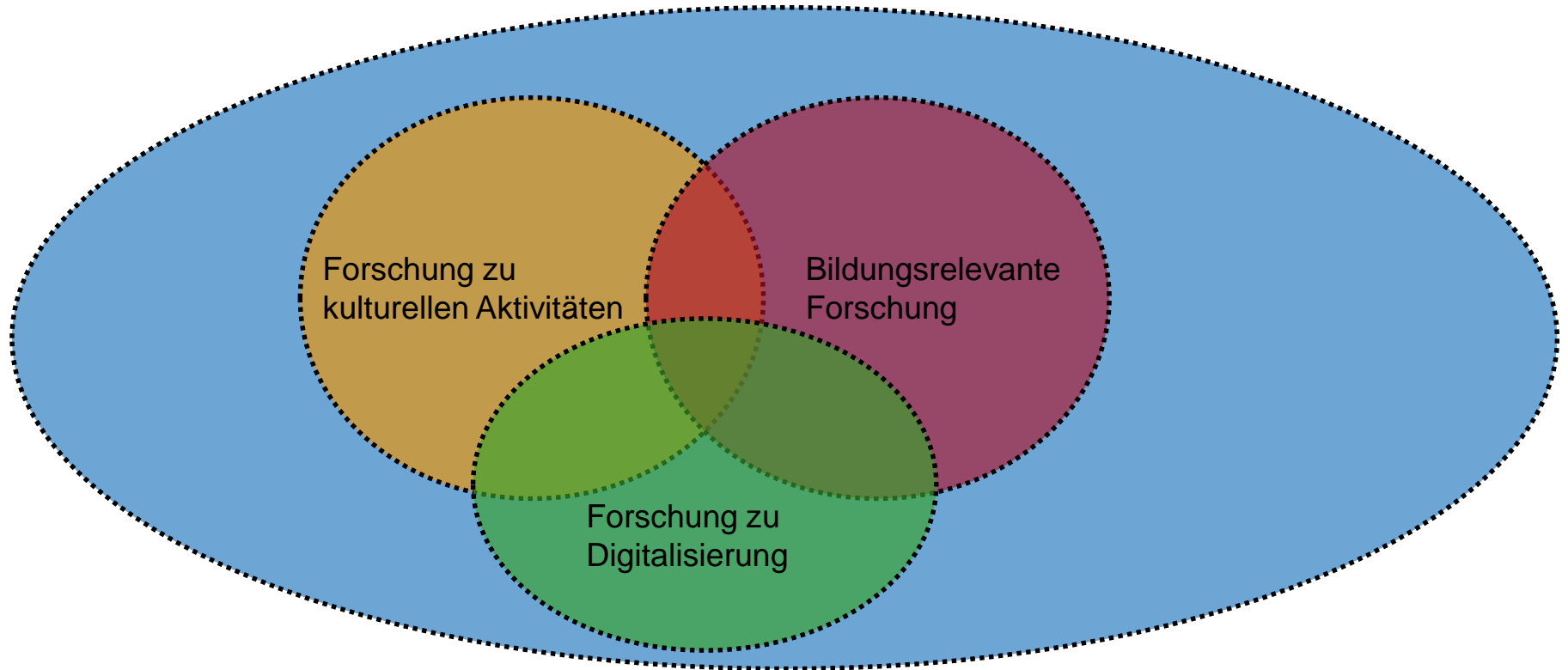
Digitalisierung in der kulturellen Bildung
(DiKuBi)

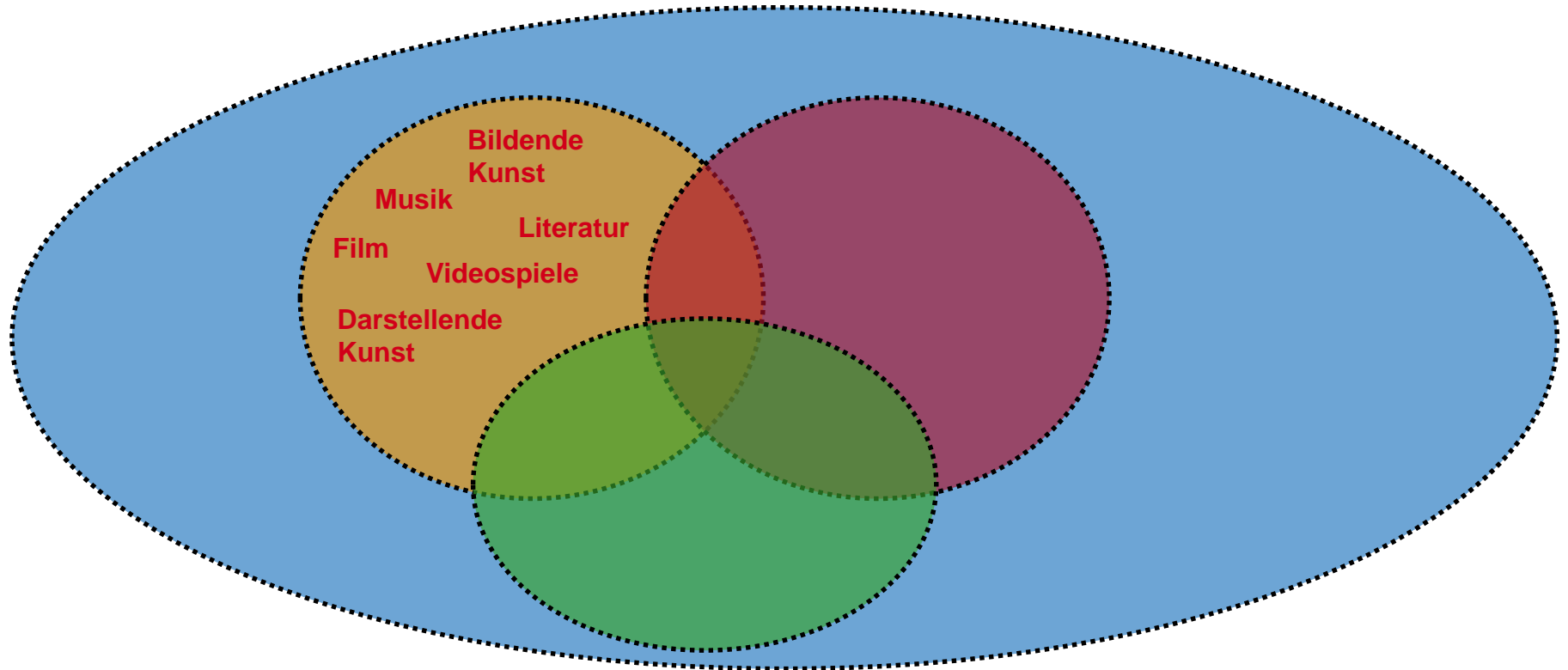


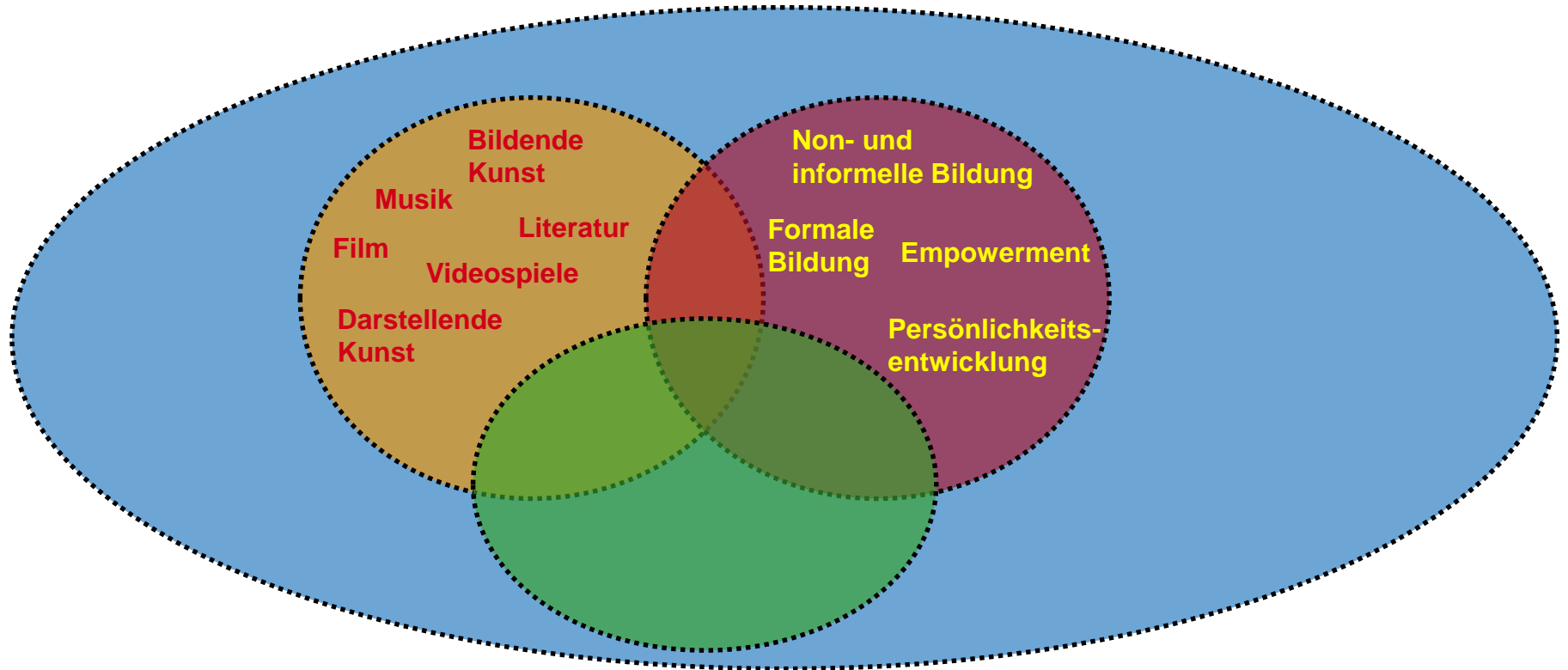


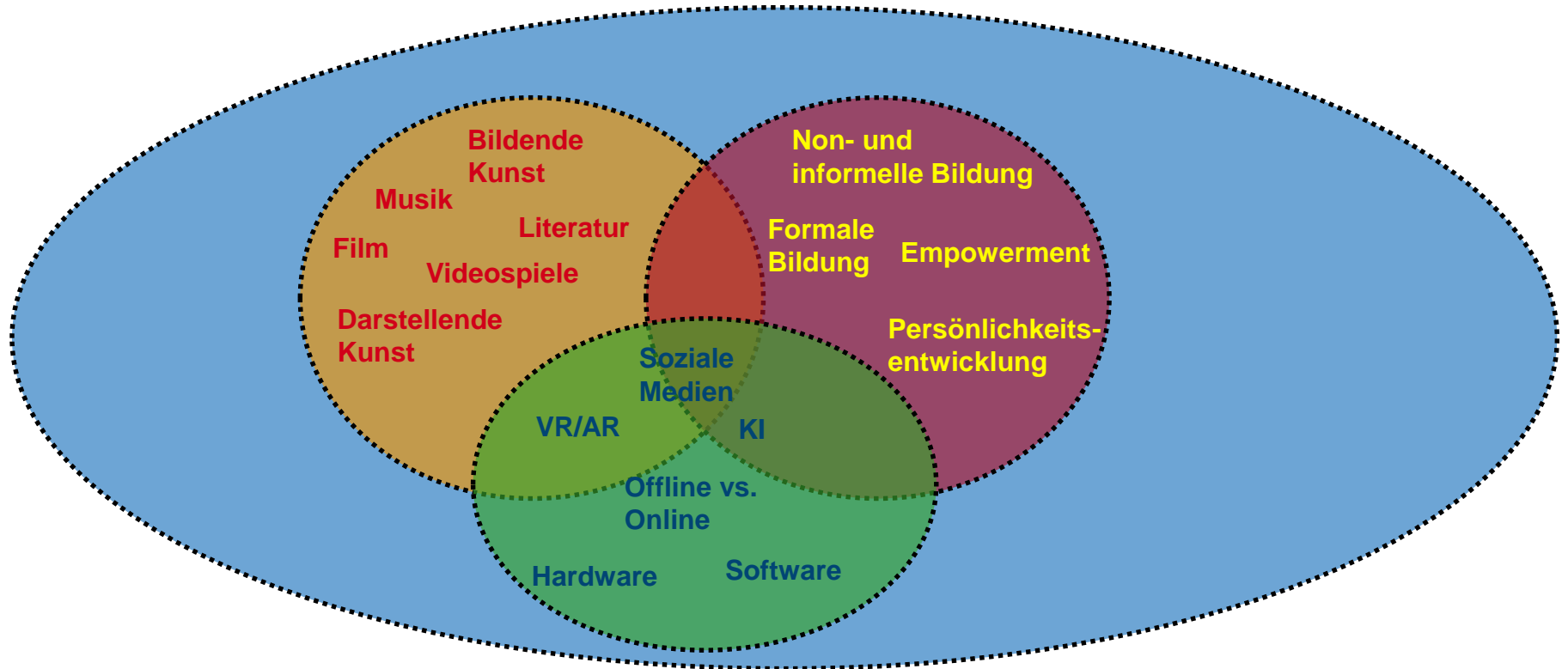
Forschung zu
kulturellen Aktivitäten



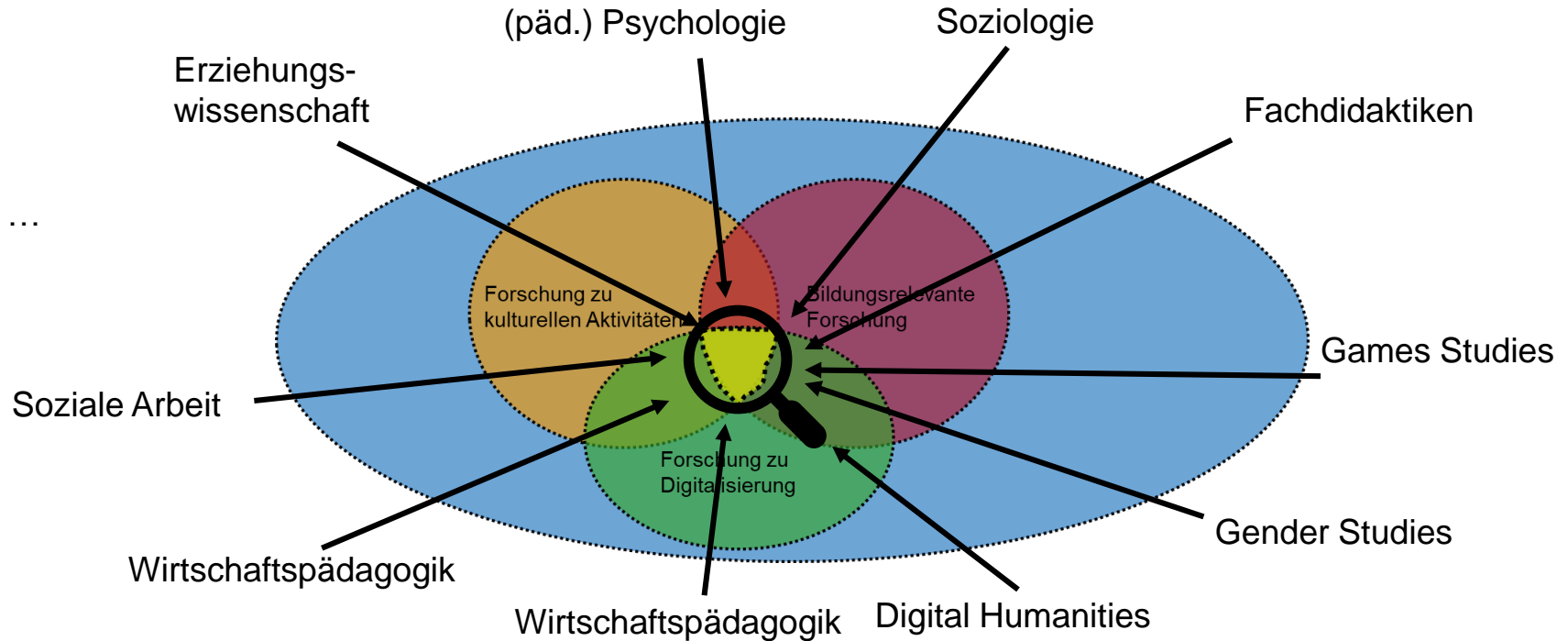


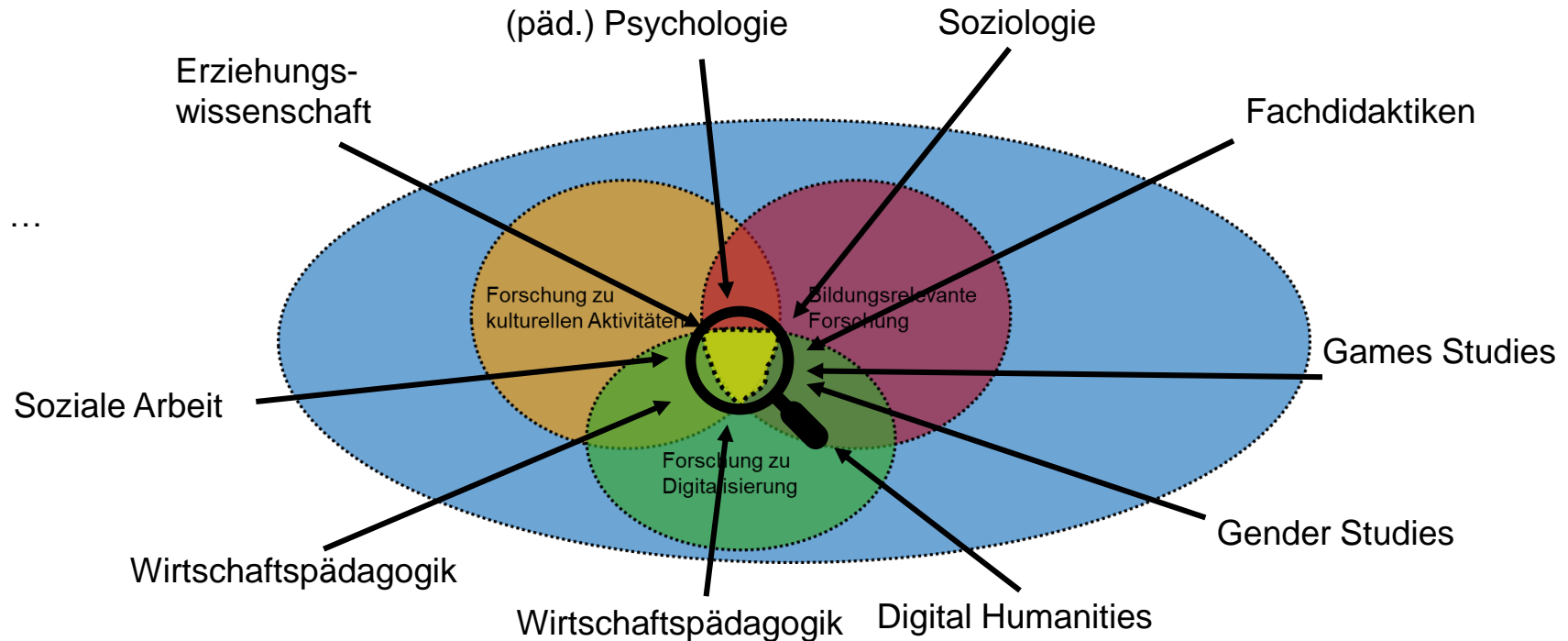










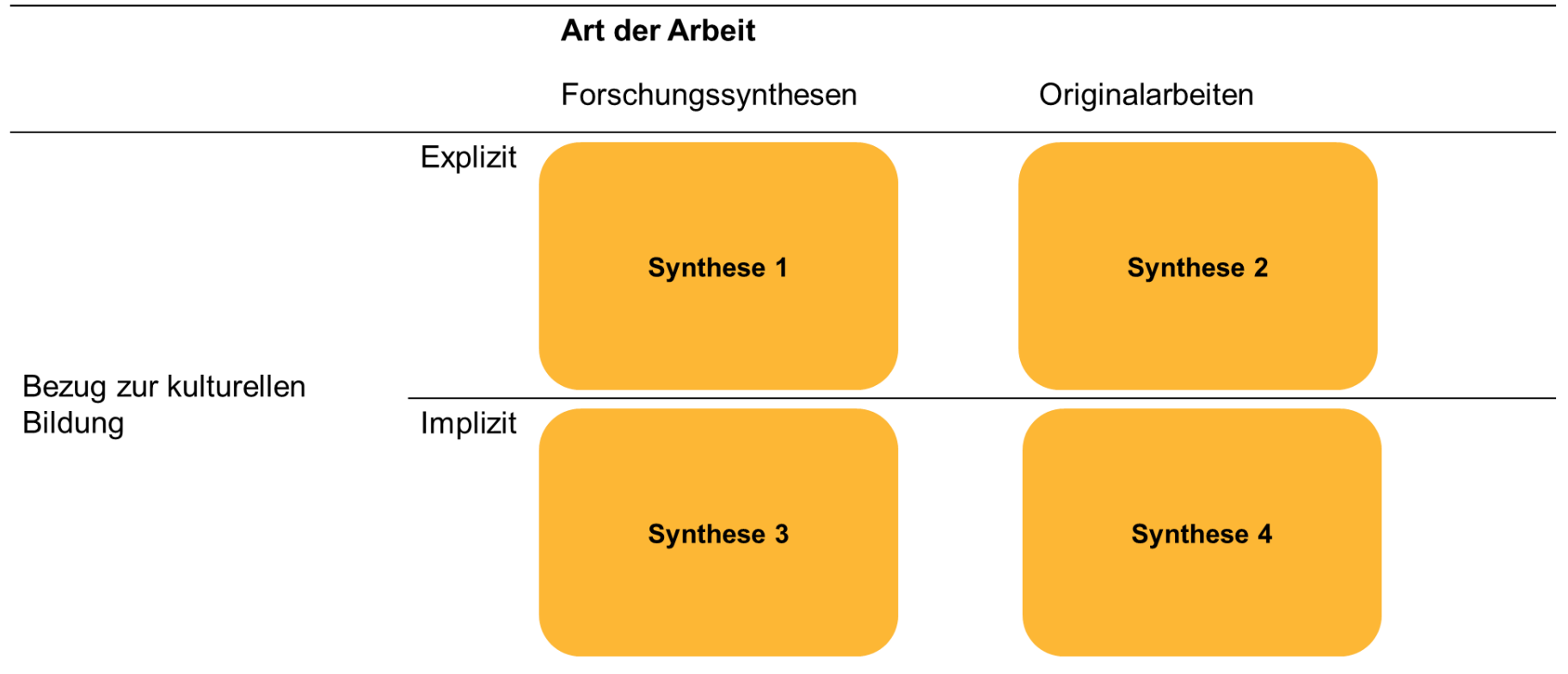


- Disziplinspezifische Fragestellungen, methodische und theoretische Paradigmen und eigene Terminologien erschweren Transfer der Ergebnisse zwischen den Disziplinen und die Aufarbeitung des Forschungsfeldes

-
- Keine Kenntnisse über:
 - vorhandene Forschungssynthesen
 - Schwerpunkte
 - Trends
 - Desiderate

 - Aufarbeitung der internationalen Forschung zur DiKuBi mit mehreren Forschungssynthesen

 - Strukturierung der internationalen Forschung nach:
 - (1) Bezug zur kulturellen Bildung und
 - (2) Art der Arbeit



Art der Arbeit

Forschungssynthesen

Originalarbeiten

Explizit

Bezug zur kulturellen
Bildung

Implizit

**Schwerpunkthemen der quantitativ-empirischen
Forschung mit Bezug zur Digitalisierung in der
kulturellen Bildung: Eine kartierende
Forschungssynthese
(Christ et al., 2024)**

Datamining für die Literaturrecherche

Precision and Sensitivity

Relevante Arbeit in Datenbank

1 := Ja

0 := Nein

1 := Ja

True Positives

False Positives

Mit Suchbefehl
gefunden

0 := Nein

False Negatives

True Negatives

Relevante Arbeit in Datenbank

1 := Ja

0 := Nein

1 := Ja

True Positives

False Positives

Mit Suchbefehl
gefunden

0 := Nein

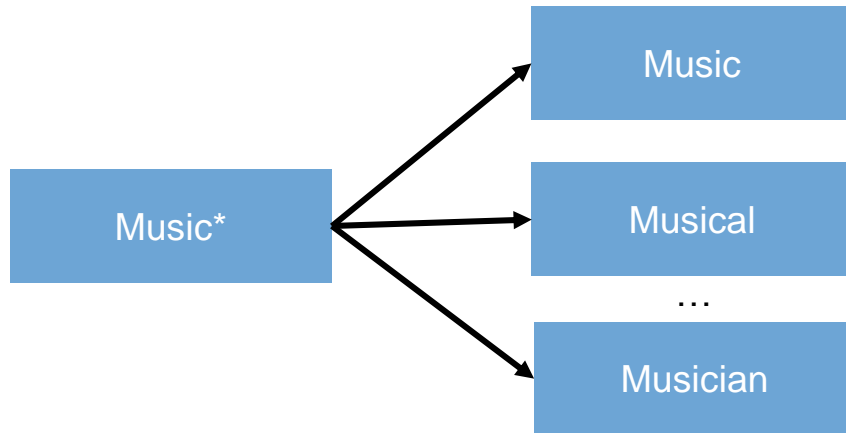
False Negatives

True Negatives

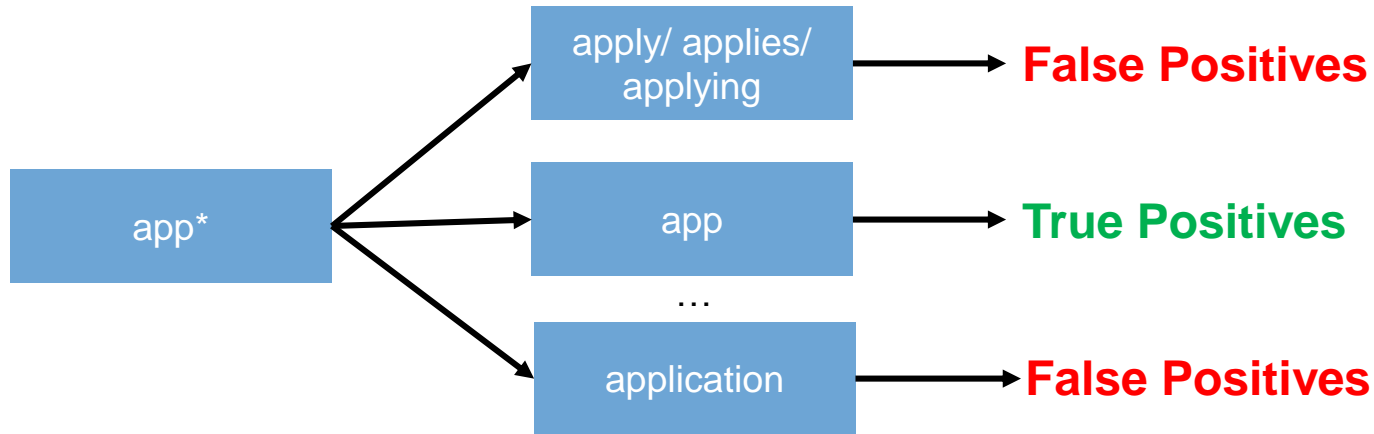
Anwendung eines breiten Suchbefehls aus drei Dimensionen (verknüpft mit AND; Begriffe verknüpft mit OR):

- Digitalisierung/Digitalität/Digitale Medien
(z.B. digital*, media*, app*, computer, internet, online...)
 - Kulturellen Aktivitäten
(z.B. cultur*, music*, instrument*, art*, videogam*, literatur*...)
 - Bildung
(z.B. educat*, learn*, cognit*, efficac* ...)
- Insgesamt n = 752 888 Datenbanktreffer (für 2000 – 2020)

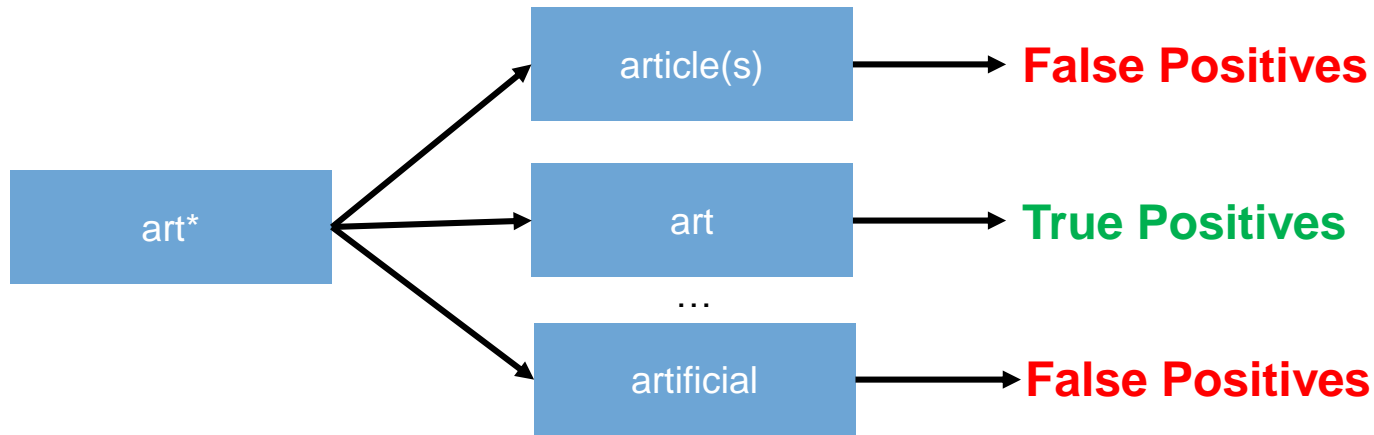
- Für alle verwendeten Wortstämme wurden ihre Grundformen ausgegeben:



- Wortstämme, die zu vielen False Positives geführt haben, wurden durch alle relevanten Permutationen ersetzt.



- Wortstämme, die zu vielen False Positives geführt haben, wurden durch alle relevanten Permutationen ersetzt.



- Analyse, welche der Suchbegriffe gemeinsam auftreten
- Ausgabe an Dokumentenlisten, die innerhalb einer Dimension (z.B. Musik oder Digitalisierung) nur „randständige“ Suchbegriffe beinhalten (z.B. media, instrument)
- Sichtung der Listen
- Möglichkeiten:
 - (a) Löschung von Suchbegriffen, die nur zu False Positives geführt haben (z.B. Instrument)

oder

 - (b) „Erweiterung“ der Suchbegriffe um False Positives zu reduzieren (z.B. media* -> „digital media“/ “social media“)

Überarbeitung Suchbefehl 3: Erweiterung um häufige und indikative Wörter

- Analyse nach relevanten Wörtern, um False Negative Rate zu reduzieren lieferte weitere Suchbefehle wie:
 - Pokémon
 - Avatar
 - Meme
 - SecondLife
 - Adventure
 - Selfie
 - Flow
 - Kinect
 - Immersion

Suchbefehl	Anzahl Treffer
Initialer Suchbefehl	752 888
Ersetzung app*	434 389
Ersetzung art*	314 721
Ersetzung act*	215 234
Löschung/Erweiterung Suchbefehl	168 264
Hinzufügung weiterer Begriffe	214 638

Aufbereitung Korpus

Bereinigung und Scoring

Bereinigung aller Texte durch:

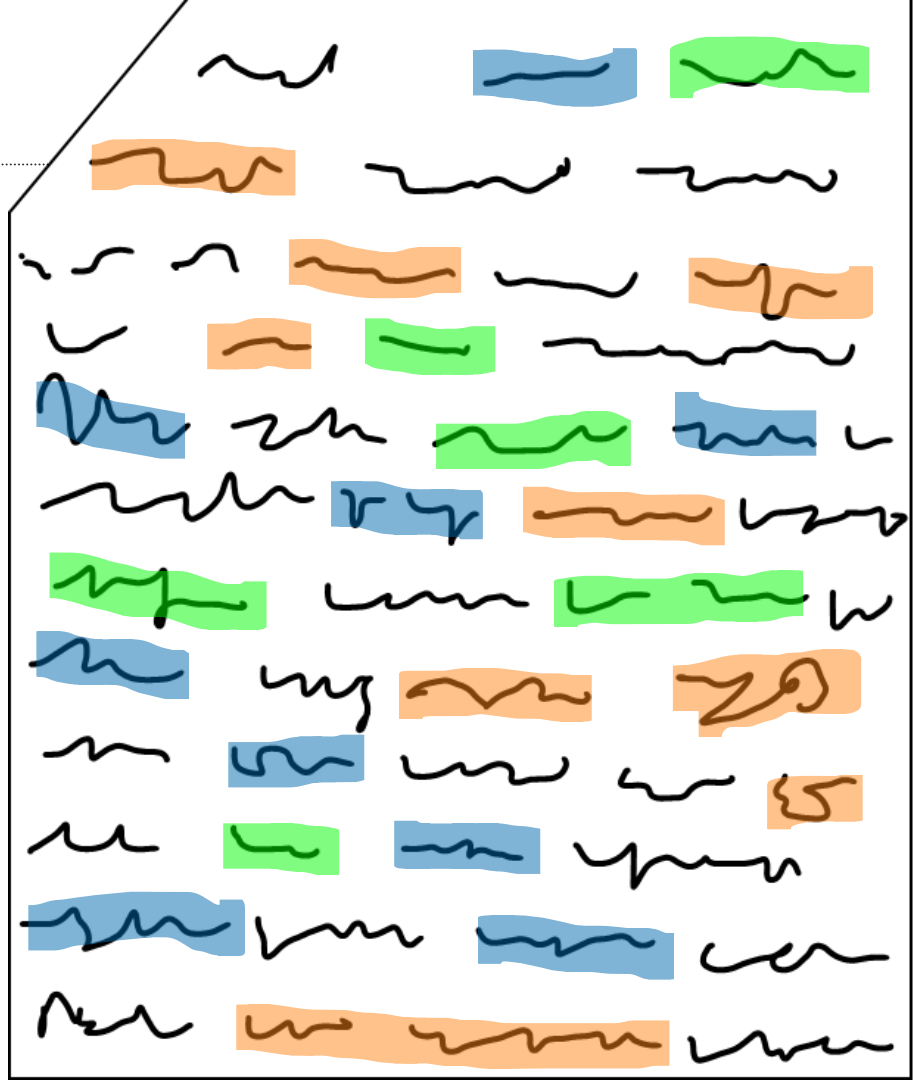
- Entfernung von Stop Words
(z.B. Bindewörter)
- Entfernung irrelevanter Textteile
(z.B. Abstracts/Titel in Originalsprache, Copyright-Statements)
- Reduktion aller Wörter auf Wortstämme
- Substitution bestimmter Wortkombinationen
(z.B. game theory -> game_theory, social media -> social_media)

Scoring der Texte

Jeder Text (Titel, Abstract, Keywords, Journal) wurde gescored nach dem Anteil der Wörter, die indikativ für die Dimensionen von DiKuBi waren.

Daraus resultierten Scores für:

- **Digitalisierung**
- **Kulturelle Aktivitäten (z.B. Musik, Literatur)**
- **Bildung**
- **Negativer Score**
(mit Termen wie „medical“, „hiv“, „game theory“, „aphasia“, „anti-retroviral therapy“)



- Auflistung aller Journals mit substantiellen, mittleren negativen Signifikanzwerten
 - Sichtung der Journaltitel und ggf. Beschreibung der Journals
 - Ausschluss irrelevanter Journals

- Auflistung aller Journals mit substantiellen, mittleren negativen Signifikanzwerten
- Sichtung der Journaltitel und ggf. Beschreibung der Journals
- Ausschluss irrelevanter Journals

Suchbefehl	Anzahl Treffer
Initialer Suchbefehl	752 888
Ersetzung app*	434 389
Ersetzung art*	314 721
Ersetzung act*	215 234
Löschung/Erweiterung Suchbefehl	168 264
Hinzufügung weiterer Begriffe	214 638
Ausschluss irrelevanter Journals	189 772

- Auflistung aller Journals mit substantiellen, mittleren negativen Signifikanzwerten
 - Sichtung der Journaltitel und ggf. Beschreibung der Journals
 - Ausschluss irrelevanter Journals

- Verwendung der Scores für die Identifikation relevanter Arbeiten

Identifikation relevanter Arbeiten

Predictive Regression Modelling, Support
Vector Machines und Neuronale Netzwerke

„The [no free lunch] theorem states that all optimization algorithms perform equally well when their performance is averaged across all possible problems.”

- Jason Brownlee, angelehnt an Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, 1(1), 67-82. [10.1109/4235.585893](https://doi.org/10.1109/4235.585893)

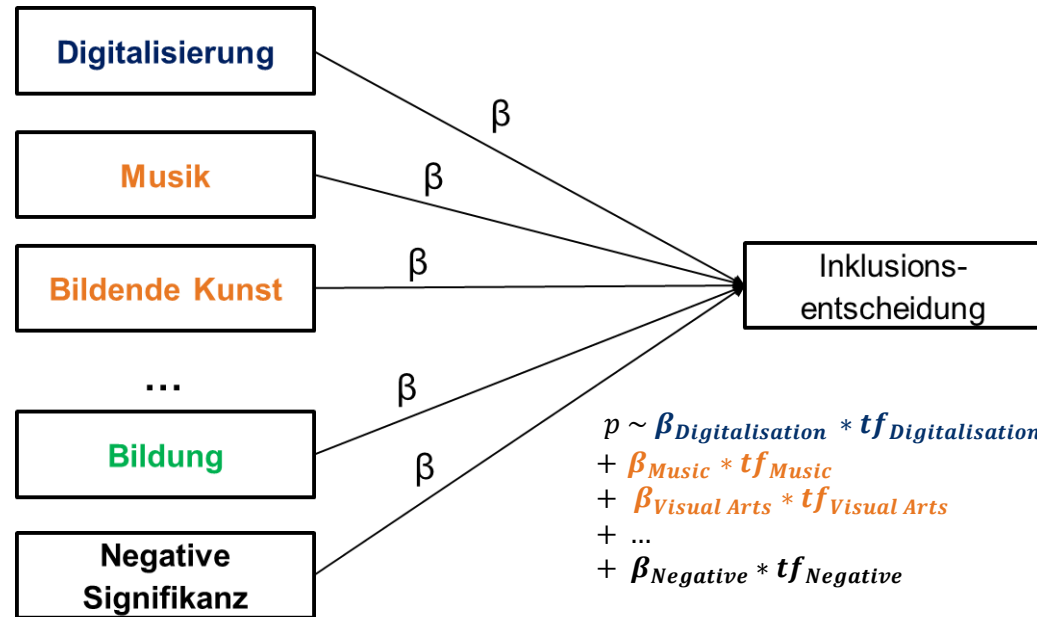
[<https://machinelearningmastery.com/no-free-lunch-theorem-for-machine-learning/>]

Verfahren	Eignung	Vor- und Nachteile
„klassische“ Ansätze		
Gewichtung nach Keywords	Enge Forschungsfrage mit wenigen, sehr wichtigen Keywords	+ Einfache Anwendung klassischer statistischer Methoden
Predictive Regression Modelling	Breite Forschungsfrage mit großer Anzahl potentiell relevanter Keywords	- Bag-of-Words-Ansatz
Simulationsansätze/Black-Box-Systeme		
Topic Modelling	Korpora mit heterogenen, klar abgrenzbaren Themen	+ Ausschluss irrelevanter Dokumentencluster - Qualitative Analyse der Cluster zeitaufwändig
Support Vector Machines	Korpora mit „wenigen“ Dimensionen	„Curse of Dimensionality“
Neural Networks / Deep Learning	Generelle Eignung	Komplexe Black-Box-Systeme mit starken Abhängigkeiten von Hyperparametern
LLM		verbunden mit hohem Coding-Aufwand und Rechenleistung

Verfahren	Eignung	Vor- und Nachteile
„klassische“ Ansätze		
Gewichtung nach Keywords	Enge Forschungsfrage mit wenigen, sehr wichtigen Keywords	+ Einfache Anwendung klassischer statistischer Methoden
Predictive Regression Modelling	Breite Forschungsfrage mit großer Anzahl potentiell relevanter Keywords	- Bag-of-Words-Ansatz
Simulationsansätze/Black-Box-Systeme		
Topic Modelling	Korpora mit heterogenen, klar abgrenzbaren Themen	+ Ausschluss irrelevanter Dokumentencluster - Qualitative Analyse der Cluster zeitaufwändig
Support Vector Machines	Korpora mit „wenigen“ Dimensionen	„Curse of Dimensionality“
Neural Networks / Deep Learning	Generelle Eignung	Komplexe Black-Box-Systeme mit starken Abhängigkeiten von Hyperparametern verbunden mit hohem Coding-Aufwand und Rechenleistung
LLM		

Nutzung der Signifikanzscores zur Identifikation relevanter Arbeiten in einem iterativen Machine-Learning-Verfahren:

1. Erstellung Training Set
 2. Erklärung der Varianz der Sichtungentscheidung im Training Set mit Signifikanzscores
 3. Vorhersage der Inklusionswahrscheinlichkeit p im Test Set
 4. Sichtung der Arbeiten mit hoher Inklusionswahrscheinlichkeit
- Gesichtete Arbeiten werden Training Set hinzugefügt und nächste Iteration startet



- Training Set aus vorherigen Synthesen
- Bei jeder Iteration wurden die top $n = 500$ Arbeiten gesichtet

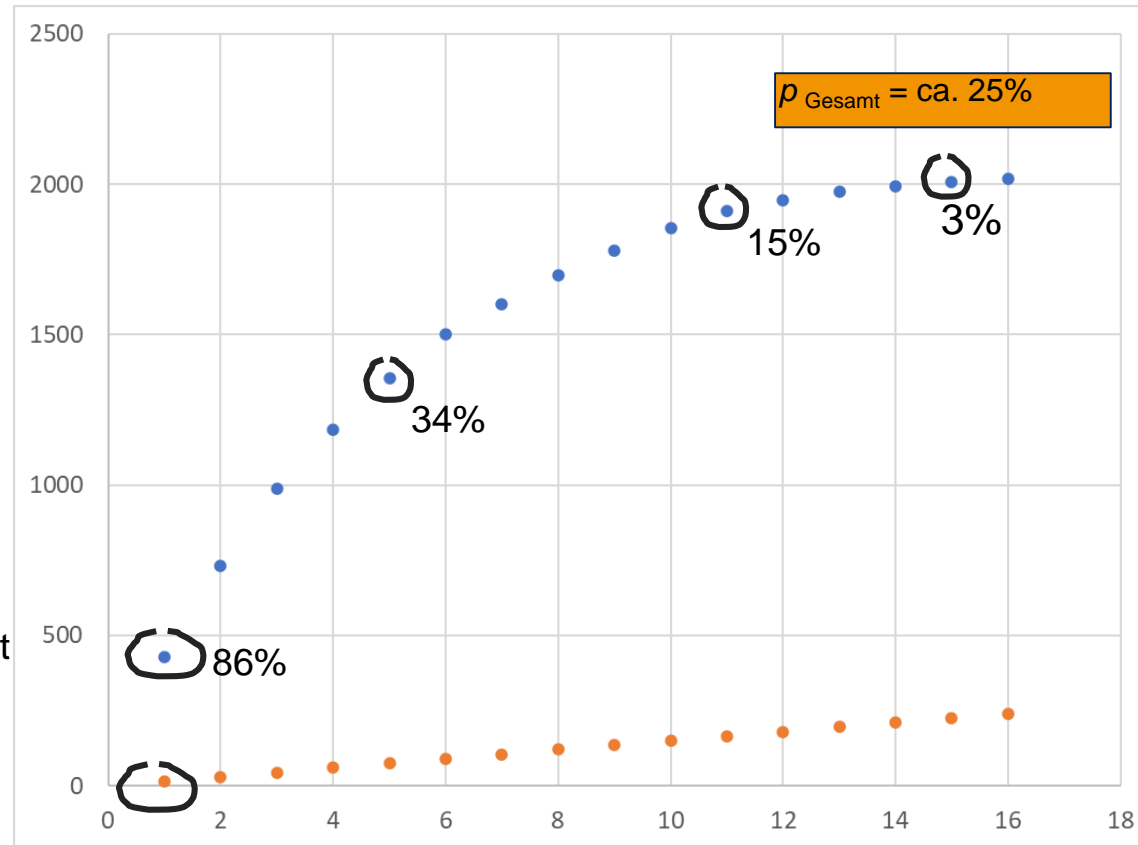
Iteration 1:

$p = .86$ im Vgl. zu $p_{\text{Zufall}} = 0.023$

Iterationen 2 – 16:

sinkende Inklusionswahrscheinlichkeit,
stabile Varianzaufklärung von
 $R^2 \in [.46; .52]$

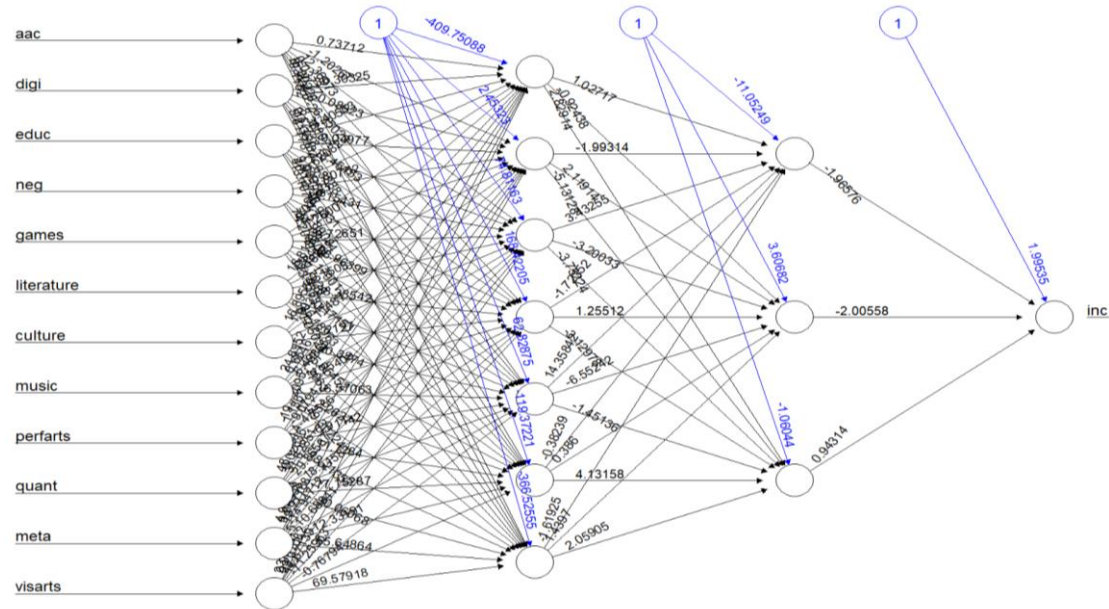
Abbruch nach Iteration 16, da
vorhergesagte Inklusionswahrscheinlichkeit
nahe 0



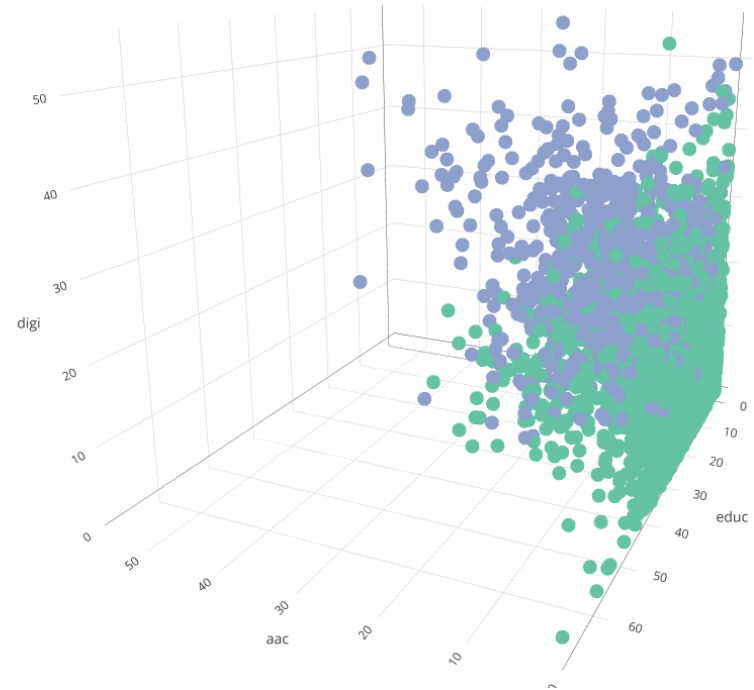
- Exploratives Verfahren zur Bestimmung eines komplexen Netzwerks aus Gewichten, die zur Einschätzung der Relevanz von Dokumenten führen können.

Vorgehen:

1. Berechnung des NN und der Gewichte im Training Set.
2. Anwendung des resultierenden NN im Test Set.



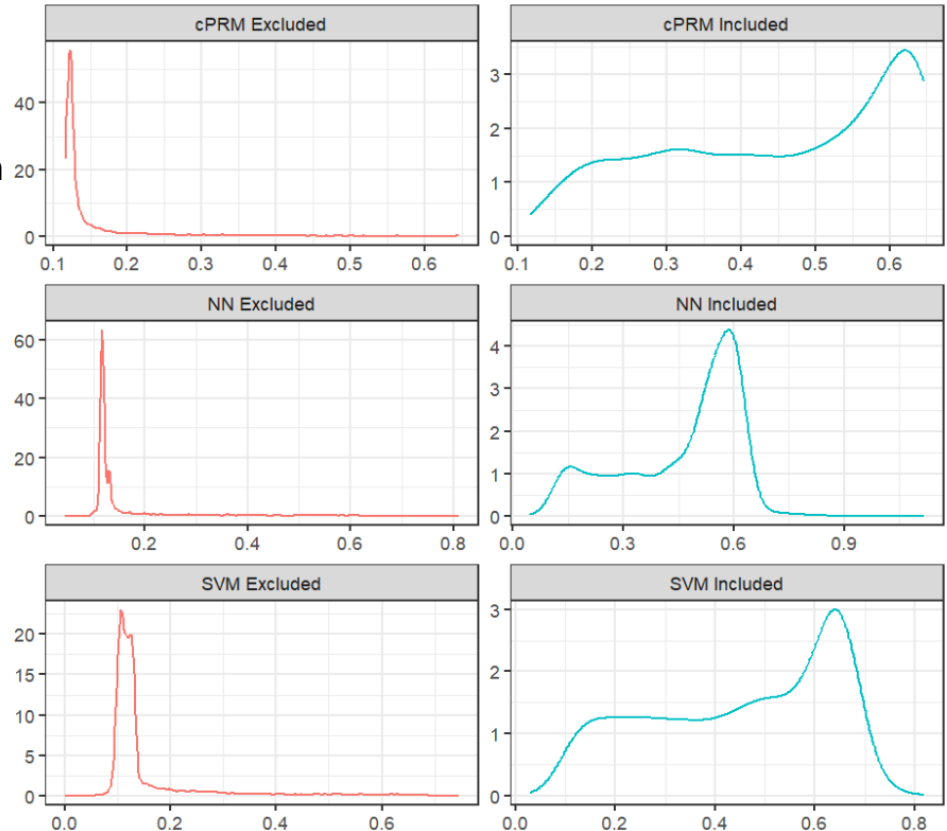
- Annahme: Die Dokumente bzw. die ihnen zugewiesenen Variablen spannen einen k-dimensionalen Raum auf. In diesem Raum können Hyperebenen bestimmt werden, die die Dokumente nach relevant vs. nicht-relevant clustern.
- Vorgehen:
 1. Bestimmung der Hyperebenen im Training Set.



- Annahme: Die Dokumente bzw. die ihnen zugewiesenen Variablen spannen einen k -dimensionalen Raum auf. In diesem Raum können Hyperebenen bestimmt werden, die die Dokumente nach relevant vs. nicht-relevant Clustern.
- Vorgehen:
 1. Bestimmung der Hyperebenen im Training Set.
 2. Unterteilung der Dokumente im Test Set mit den Hyperebenen.



- Ähnliche Ergebnisse für alle drei Verfahren.
- Hohe Korrelationen zwischen allen drei vorhergesagten Inklusionswahrscheinlichkeiten ($r \in [.90; .92]$)
- Ranking Aufwand:
NN > SVM > Predictive Regression Modelling



Analyse und Kategorisierung

Beispiel: Topic Modelling

Topic modelling ist ein statistisches (fuzzy) Clustering-Verfahren zur Identifikation latenter Themen (sog. Topics).

Annahmen:

- Im Korpus werden k Themen behandelt.
- Die Themen können durch die Wörter bestimmt werden, die in den Dokumenten (gemeinsam) vorkommen.
- Jedem Wort kann pro Topic ein Relevanzgewicht β zugeordnet werden.
- Jedem Dokument kann pro Topic ein Relevanzgewicht γ zugeordnet werden.

Die Anzahl der Topics k wird per Gibbs-Sampling und qualitativer Beurteilung der Topic(-Cluster) bestimmt.

In diesem Beispiel wurden 2 Topic Models bestimmt: (1) kulturelle Aktivitäten und (2) latenten Themen jenseits der kulturellen Aktivitäten.

Topic Modelling ergab $k = 17$ Themen bzgl. Untersucher kultureller Aktivitäten und $k = 22$ übergreifende latent Themen.

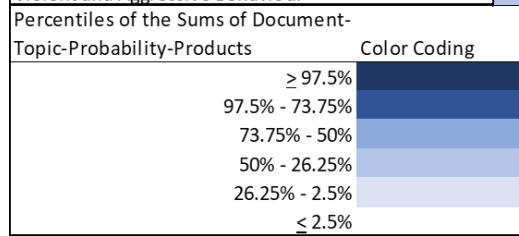
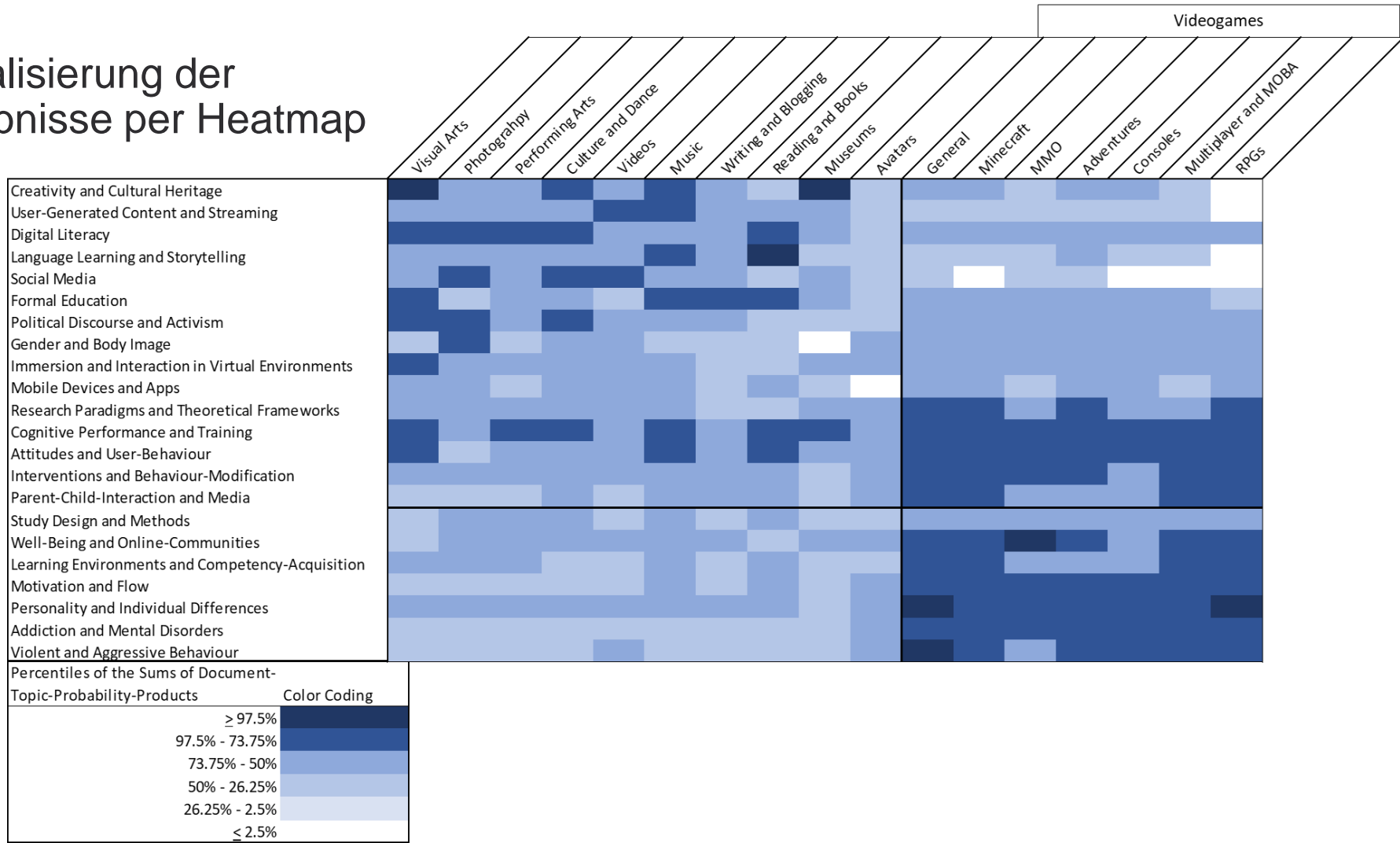
Kulturelle Aktivitäten:

- Jede Facette kultureller Aktivität wurde untersucht (d.h. Musik, Literatur, darstellende Kunst, bildende Kunst). Eindeutiger Schwerpunkt Videospiele mit $k = 7$ Themen.

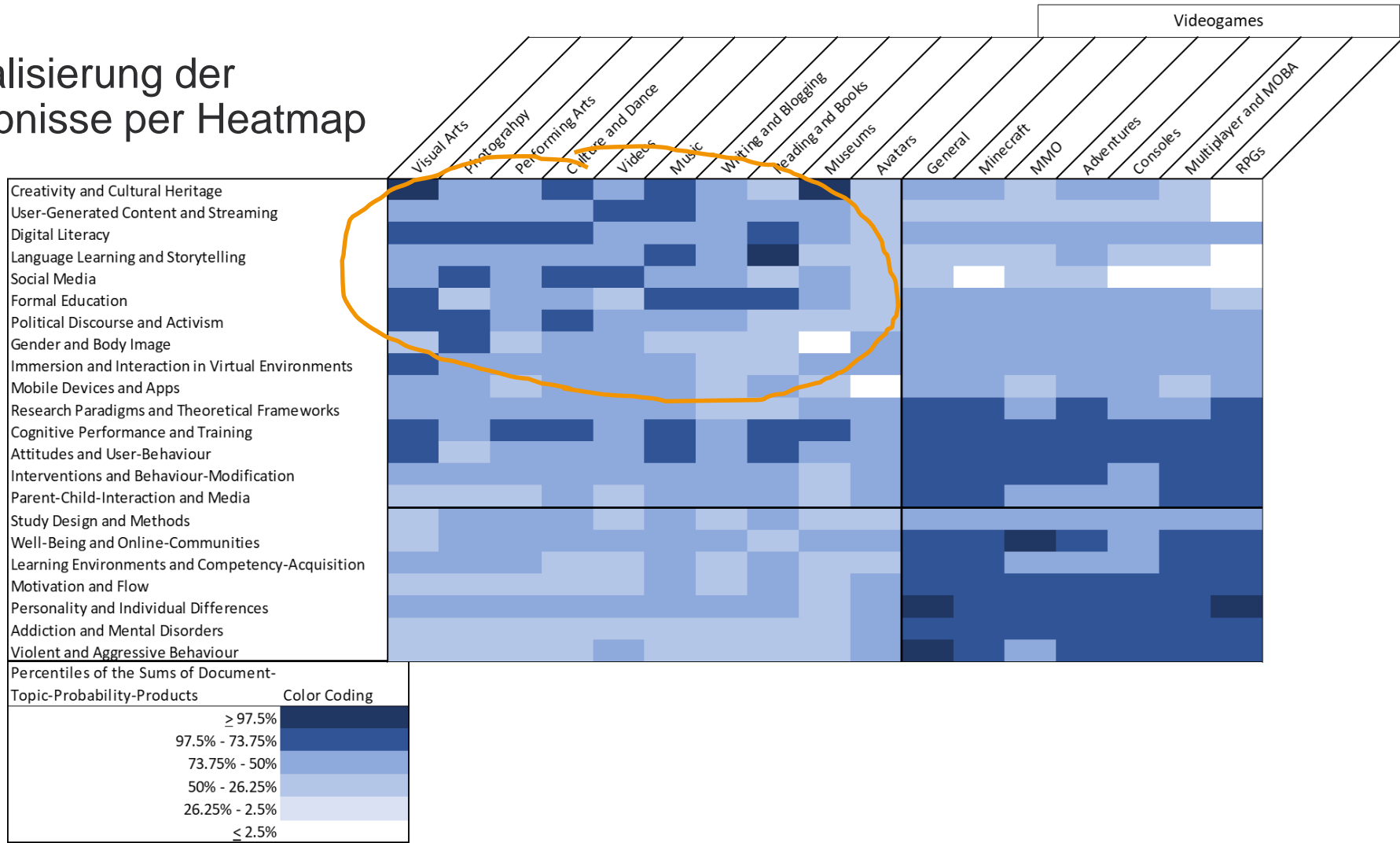
Übergreifende Themen:

- Übergreifende Themen waren sehr heterogen mit z.B.:
 - “creativity and cultural heritage”
 - “formal education”
 - “aggressive and violent behaviour”

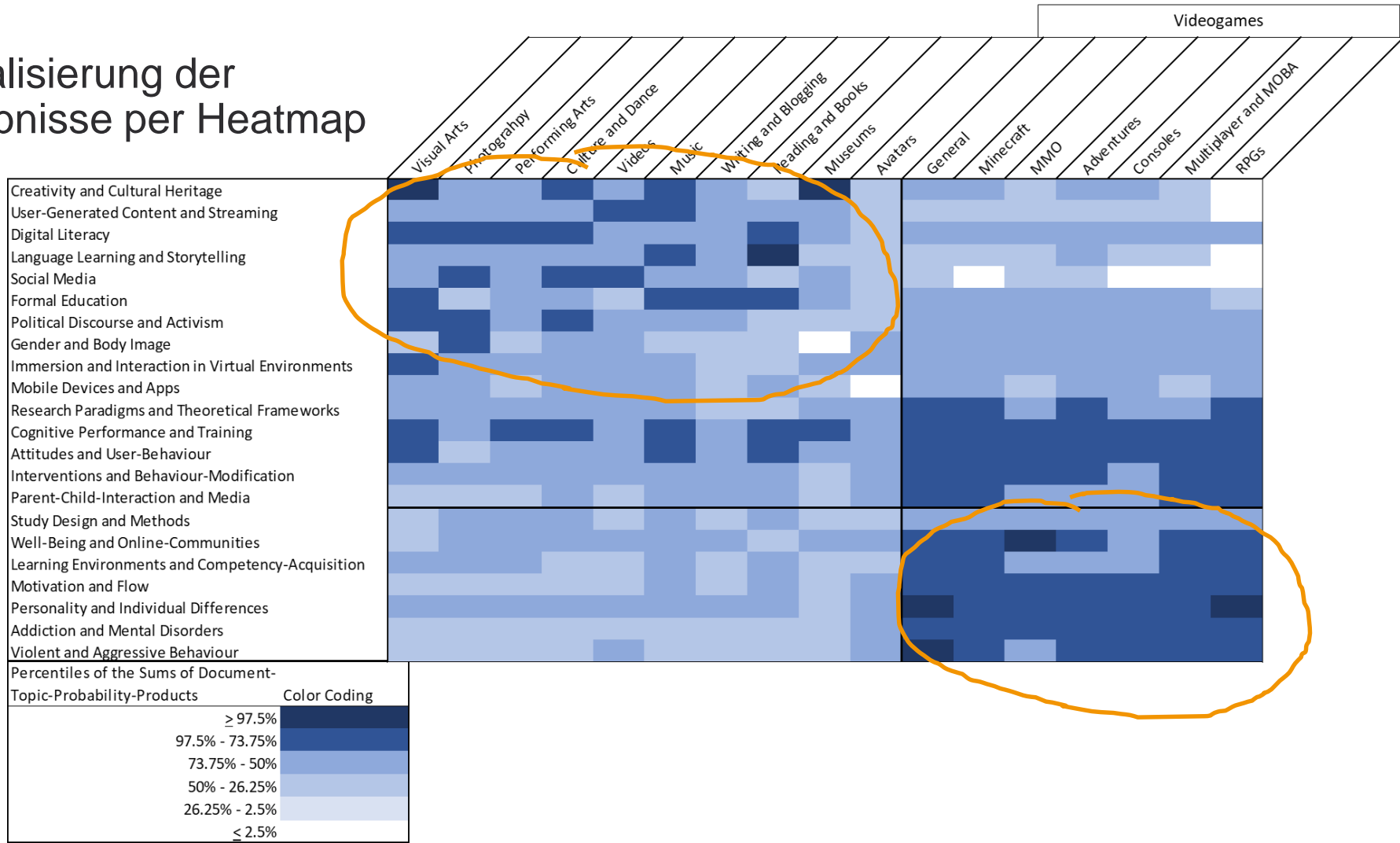
Visualisierung der Ergebnisse per Heatmap



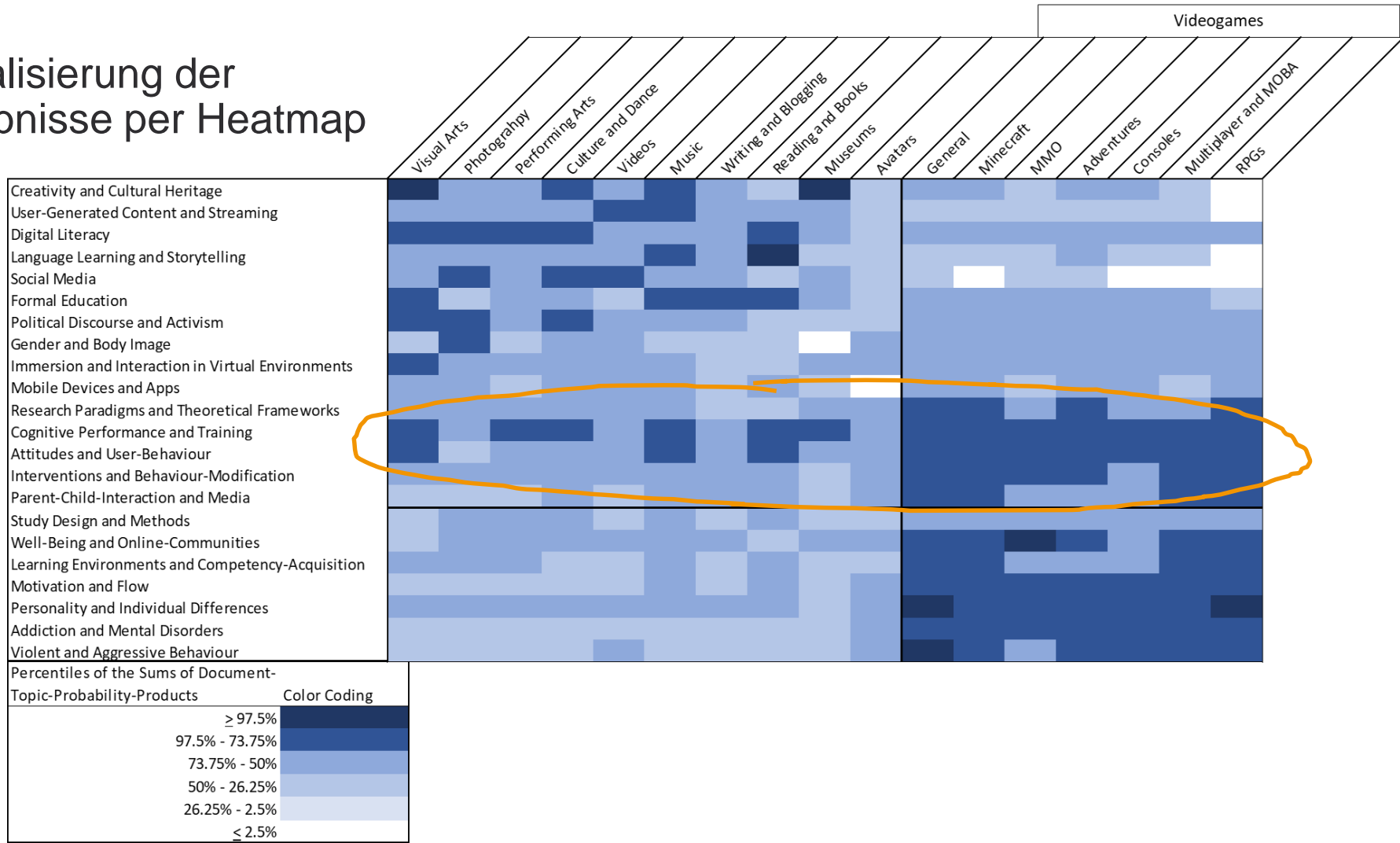
Visualisierung der Ergebnisse per Heatmap



Visualisierung der Ergebnisse per Heatmap



Visualisierung der Ergebnisse per Heatmap



Fazit und Ausblick

-
- Data Mining eignet sich gut um große Literaturkorpora zu sortieren, potentiell relevante Arbeiten zu identifizieren und liefert dadurch eine erhebliche Zeitersparnis
 - Machine Learning Verfahren muss je nach Breite der Fragestellung und Heterogenität des Forschungsgegenstands ausgewählt werden. Häufig kommt man dabei um Ausprobieren nicht herum.
 - Verwendung von Machine Learning führt bei Forschungssynthesen immer zum Übersehen potentiell relevanter Arbeiten. Die resultierende False Negative Rate kann jedoch quantifiziert werden.
- Welches Verfahren sich am besten eignet, lässt sich mit einem repräsentativen Training Set am besten identifizieren!

- Vergleich der Ergebnisse der vorgestellten Verfahren mit „modernen“ AI-Tools steht noch aus
- Integration weiterer Kovariaten (z. B. Affiliation, Zitationsnetzwerke) und Variablen-Clustern (z. B. Autor:in, Arbeitsgruppe, Journal)
- Einstiegshürde für Verfahren verringern durch Entwicklung von Tools oder Workshops

Vielen Dank für die Aufmerksamkeit!

Ich freue mich auf Fragen und Anregungen.


Schwerpunktt Themen der quantitativ-empirischen Forschung mit Bezug zur Digitalisierung in der kulturellen Bildung: Eine kartierende Forschungssynthese


Hot topics of quantitative-empirical research related to digitalization in cultural education: a mapping review

Allgemeiner Teil | [Open access](#) | Published: 04 December 2023

Volume 27, pages 351–392, (2024) | [Cite this article](#)

[Download PDF](#) 

 You have full access to this [open access](#) article

[Alexander Christ](#) , [Kathrin Smolarczyk](#) & [Stephan Kröner](#)

<https://doi.org/10.1007/s11618-023-01210-7>

