Swiss Journal of Sociology, 51 (2), 2025, 337–359

Metabilder als Forschungswerkzeuge. Zur Kontingenz und algorithmischen Bedingtheit ihrer Herstellung

Sebastian W. Hoggenmüller* und Harald Klinke**

Zusammenfassung: Metabilder sind ein zentrales Forschungswerkzeug in der computergestützten Analyse großer visueller Datenbestände. Basierend auf algorithmischen Verfahren machen sie Muster, Anomalien und andere signifikante Zusammenhänge in den Daten sichtbar, die andernfalls möglicherweise unentdeckt blieben. Der Beitrag nimmt Metabilder als Ergebnisse eines kontingenten, entscheidungsabhängigen Prozesses in den Blick und hinterfragt sie aus einer kritischen Perspektive, die Visuelle Soziologie und Digitale Bildwissenschaft miteinander verbindet.

Schlüsselwörter: Algorithmische Datenvisualisierung, computergestützte Bildanalyse, Image Plot, Visuelle Soziologie, Digitale Bildwissenschaft

Metapictures as Research Tools. On the Contingency and Algorithmic Conditionality of Their Production

Abstract: Metapictures are a key research tool in the computational analysis of large visual datasets. Based on algorithmic processes, they reveal patterns, anomalies, and other significant relationships within the data that might otherwise remain undiscovered. This article considers metapictures as results of a contingent, decision-dependent process, and investigates them from a critical perspective that combines visual sociology and digital image science.

Keywords: Algorithmic data visualisation, computational image analysis, image plot, visual sociology, digital image science

Les métapictures en tant qu'outils de recherche : contingence et la conditionnalité algorithmique de leur production

Résumé : Les métapictures constituent un outil de recherche essentiel dans l'analyse assistée par ordinateur de grands ensembles de données visuelles. Basées sur des processus algorithmiques, elles permettent de révéler des schémas, des anomalies et d'autres liens significatifs au sein des données qui resteraient autrement potentiellement inaperçus. Cet article examine les métapictures comme le résultat d'un processus contingent et tributaire de décisions, et les interroge dans une perspective critique croisant la sociologie visuelle et la science de l'image numérique.

Mots-clés : Visualisation algorithmique des données, analyse d'images assistée par ordinateur, image plot, sociologie visuelle, science de l'image numérique

* Universität Luzern, Kultur- und Sozialwissenschaftliche Fakultät, Soziologisches Seminar, CH-6002 Luzern, sebastian.hoggenmueller@unilu.ch.

** LMU München, Institut für Kunstgeschichte, D-80798 München, h.klinke@lmu.de.

DOI 10.26034/cm.sjs.2025.6933

1 Einleitung – von großen visuellen Datenbeständen zu visuellen Mustern[1]

Die jüngsten Fortschritte in der Bildverarbeitung, im maschinellen Sehen (Computer Vision) und in der Künstlichen Intelligenz haben der kultur-, geistes- und sozialwissenschaftlichen Erforschung großer visueller Datenbestände entscheidende Impulse gegeben. Ein zentrales Forschungswerkzeug, das auf diesen Entwicklungen beruht, sind sogenannte Metabilder – besser bekannt als Image Plots, ein Begriff, der maßgeblich von dem Medientheoretiker und Künstler Lev Manovich (vgl. vor allem 2012a; 2020) etabliert wurde. Dabei handelt es sich um computergestützte und algorithmusbasierte Visualisierungen großer Mengen digitaler visueller Daten, die es erlauben, Muster, Anomalien, Ausreißer und andere signifikante Zusammenhänge in den Datenbeständen zu entdecken, und zwar solche Zusammenhänge, die allein schon aufgrund der enormen Datenmengen für das bloße Auge in der Regel nicht erkennbar und durch manuelle Analysen kaum zu erfassen sind. Metabilder eröffnen mithin Perspektiven auf visuelle Phänomene und deren Erforschung, die ohne sie in dieser Form nicht zugänglich wären. Beispiele für ihre Anwendung reichen von der Untersuchung künstlerischer Entwicklungen in umfangreichen Kunstsammlungen (vgl. z. B. Murphy et al., 2022; Frischknecht in diesem Sonderheft) über die Analyse der Flut visueller Inhalte auf Social-Media-Plattformen (vgl. etwa Rogers, 2021; Hochman & Schwartz, 2021) bis hin zur Erforschung der semantischen und visuellen Struktur historischer Filme sowie der länderspezifischen Unterschiede in der Gestaltung einer Vielzahl von Buchcovern (vgl. Manovich, 2012b; Jeong & Han, 2015).

Gegenstand des vorliegenden Beitrags sind Metabilder als spezifisches Forschungswerkzeug der Kultur-, Geistes- und Sozialwissenschaften und als nicht selbstverständliches Ergebnis eines vorausgegangenen Herstellungsprozesses. Drei Fragen stehen dabei im Mittelpunkt: Was genau sind Metabilder? Wie werden sie erzeugt? Und welche Herausforderungen gehen mit einer kritischen Nutzung von Metabildern als Forschungswerkzeug einher? Diesen Fragen nähern wir uns aus einer interdisziplinären Perspektive, die unsere beiden Teildisziplinen – die Visuelle Soziologie und die Digitale Bildwissenschaft – miteinander verbindet. Das Ziel unserer Überlegungen ist es, eine kritische Auseinandersetzung mit Metabildern im Allgemeinen zu fördern, indem wir sowohl die Kontingenz ihres Herstellungsprozesses als auch die Rolle algorithmischer Verfahren in diesem Prozess offenlegen. Dadurch wollen wir zu einer fundierten Grundlage für die Bewertung des epistemischen Gehalts von Metabildern beitragen.

Die folgende Argumentation gliedern wir hierfür in drei Hauptschritte: Im ersten Schritt erläutern wir einige grundlegende Merkmale von Metabildern, veranschaulichen diese anhand eines konkreten Beispiels und beleuchten die Herstellung von Metabildern als kontingenten, entscheidungsabhängigen Prozess (Abschnitt 2). Im

1 Wir danken den beiden Gutachter*innen für ihre wertvollen Hinweise und konstruktive Kritik. Ihre Anmerkungen haben wesentlich zur Weiterentwicklung dieses Textes beigetragen.

zweiten Schritt setzen wir uns mit den algorithmischen Prozessen hinter Metabildern auseinander und legen dar, wie diese Prozesse konkret funktionieren. Im Fokus stehen hier das mathematisch-statistische Konzept des Merkmalsraums (Feature Space) und das Verfahren der Dimensionsreduktion (Dimension Reduction), wobei Letzteres am Beispiel des Algorithmus t-Distributed Stochastic Neighbor Embedding (t-SNE) erläutert wird (Abschnitt 3). Im dritten Schritt weiten wir unseren Blick auf eine kritische Nutzung von Metabildern als Forschungswerkzeuge aus, die über unsere Zusammenarbeit in der Visuellen Soziologie und Digitalen Bildwissenschaft hinausreicht. Dabei zeigen wir drei zentrale Herausforderungen auf, die eine verstärkte interdisziplinäre Kooperation erforderlich machen (Abschnitt 4).

2 Metabilder – Charakteristik und Herstellung

Der Begriff des Metabilds wurde grundlegend von William J. T. Mitchell in seinem Essay *Metapictures* geprägt, der im Jahr 1994 in der einflussreichen Aufsatzsammlung *Picture Theory: Essays on Verbal and Visual Representation* erschienen ist. Laut Mitchell sind Metabilder Bilder, die eine Reflexion über die Funktion und die Bedeutung, über die Macht und das Potenzial von Bildern selbst ermöglichen: „Any picture that is used to reflect on the nature of pictures is a metapicture" (Mitchell, 1994, S. 57). Ein Metabild eröffnet für Mitchell entsprechend eine Metaebene, auf der das Bild als Instrument dient, um die Spezifik des Bildlichen und die Epistemik der Bildlichkeit zu ergründen.

Unser Verständnis von Metabildern baut auf Mitchells Definition auf. Allerdings konzentrieren wir uns auf computergestützte und algorithmusbasierte Visualisierungen, die (meist sehr) große Mengen digitaler visueller Daten synoptisch miteinander verknüpfen. Die digitalen visuellen Daten werden dabei auf einer üblicherweise zweidimensionalen Fläche so angeordnet, dass sowohl lokale als auch globale Beziehungen zwischen ihnen sichtbar werden. Dadurch bieten Metabilder die Möglichkeit, übergreifende Zusammenhänge sowie (Un-)Regelmäßigkeiten in großen visuellen Datenbeständen zu entdecken, mit dem Ziel, ein genaueres Verständnis dieser Datenbestände zu gewinnen. Im Wesentlichen sind sie somit „pictures about pictures" (Mitchell, 1994, S. 36), also – nach der Definition von Mitchell – Metabilder im besten Sinne.

Ein wesentliches Charakteristikum der von uns in den Blick genommenen computergestützten und algorithmusbasierten Metabilder besteht darin, dass die ihnen zugrunde liegenden einzelnen visuellen Datenobjekte – unabhängig davon, ob es sich um digitale bzw. digitalisierte Kunstwerke und kulturelle Artefakte, visuelle Social-Media-Inhalte wie Fotografien und Thumbnails oder wissenschaftliche Grafiken und Diagramme handelt – nicht durch abstrakte Punkte oder Symbole repräsentiert werden, sondern als Miniaturbilder auf der sogenannten Plotfläche,

also der Darstellungsfläche des Metabildes, sichtbar werden. Ein weiteres zentrales Merkmal ist die interaktive Bedienbarkeit der Metabilder, die es wiederum ermöglicht, dass Forschende die Miniaturbilder auf der Plotfläche explorativ erkunden können. So lassen sich beispielsweise durch Schwenkbewegungen verschiedene Bereiche des Metabildes untersuchen, während durch Hinein- und Herauszoomen dynamisch zwischen der Gesamtheit der Daten, spezifischen Elementen innerhalb von Datengruppen und konkreten Details einzelner Datenobjekte gewechselt werden kann (vgl. grundlegend zum Verhältnis von Gesamtsammlungs-, Multiobjekt- und Einzelobjektansichten in der Datenvisualisierung Windhager et al., 2019).[2]

Zur Veranschaulichung sei ein exemplarisches Metabild herangezogen: Das Beispiel (Abb. 1) stammt aus einem laufenden eigenen Forschungsprojekt, das Kunst, Computer Vision und Visuelle Soziologie verbindet (vgl. Hoggenmüller et al. [in Begutachtung]). Dieses transdisziplinäre Forschungsprojekt bezieht sich seinerseits auf ein künstlerisches Open-Data-Projekt[3], das mithilfe von maschinellem Lernen täglich über eine Million Bilder von mehr als 10 000 Netzkameras (z. B. Webcams, Überwachungssysteme und IP-Kameras) aus aller Welt analysiert, archiviert und in Echtzeit zur Exploration über eine Website bereitstellt.[4] Ziel des Forschungsprojekts ist es, die Beobachtung der Welt durch Netzkameras in öffentlichen und (halb) privaten Räumen zu untersuchen und zugleich neue analytische und kritische Perspektiven auf computergestützte Methoden, Algorithmen und maschinelles Lernen im Kontext von Big Visual Data zu entwickeln.[5]

Das exemplarische Metabild zeigt eine Darstellung von 40 000 zufällig ausgewählten, radial nach Farbton sowie Helligkeit sortierten Bildern, die im Rahmen des künstlerischen Open-Data-Projekts an einem Tag aus Asien erfasst wurden. Die zugrunde liegenden Bilddaten wurden hierfür zunächst automatisch analysiert, wobei eine Kombination aus Deep-Learning-Modellen (z. B. MobileNetV2; vgl. Sandler et al., 2018), klassischen Computer-Vision-Methoden (z. B. SIFT; vgl. Lowe,

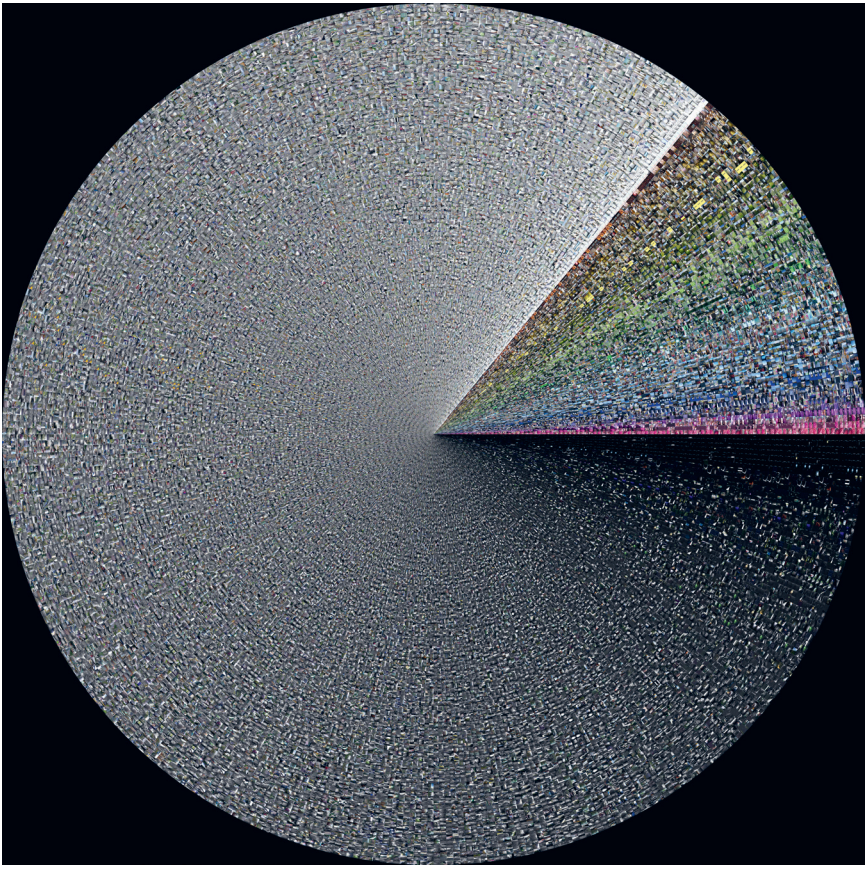
2 Eine gewisse Ähnlichkeit mit Metabildern weist das im Bereich der interpretativen Bildanalyse genutzte Forschungsinstrument der Assemblage auf, das im Rahmen der Ästhetischen Re|Konstruktionsanalyse (vgl. programmatisch Hoggenmüller, 2016, systematisch ausgearbeitet in Hoggenmüller, 2022) entwickelt wurde, um Bilddaten einzeln und vergleichend, sukzessive und simultan in den analytischen Blick zu nehmen; dabei finden zunehmend auch digitale Tools Anwendung, die durch Künstliche Intelligenz unterstützt werden (vgl. Hoggenmüller, 2025, S. 159 ff.).

3 Das Open-Data-Projekt mit dem Titel *Watching the World: The Encyclopedia of the Now* geht auf den Schweizer Fotografen Kurt Caviezel zurück, der das World Wide Web als ein Kamerasystem begreift, in dem Netzwerkkameras die Funktion der Objektive übernehmen, der Bildschirm zum Sucher wird und die Maus als Auslöser dient. Die Weiterführung des Projekts erfolgt inzwischen in Kooperation mit der Zürcher Hochschule für Angewandte Wissenschaften (für einen Einblick in das Bildarchiv des Projekts siehe Caviezel, 2015; das Projekt selbst findet sich unter: https://webcamaze.engineering.zhaw.ch/).

4 Die Netzkameras liefern via öffentlich zugänglichen URLs etwa alle zehn Minuten ein Bild, das für 48 Stunden archiviert wird, bevor es durch neuere Bilder überschrieben wird.

5 Auf personeller Ebene arbeiten in diesem Projekt Kurt Caviezel (Kunst), Fitim Abdullahu und Helmut Grabner (Computer Vision) sowie Sebastian W. Hoggenmüller (Soziologie) zusammen.

Abbildung 1 Exemplarisches Metabild: Darstellung von 40 000 Bildern, die am 10. September 2024 erfasst wurden, ausgewählt nach kontinentalem Kamerastandort (Asien) und sortiert nach Farbton und Helligkeit (radial)



Quelle: Eigene Visualisierung [mit Python erstelltes Metabild, Grabner und Hoggenmüller].

2004) und eigens trainierten Modellen (z. B. CIR(CLIP(x)); vgl. Abdullahu & Grabner, 2024) zum Einsatz kam. Die Ergebnisse der drei Analysemethoden flossen anschließend – ebenfalls automatisiert – in eine Datenbank ein, in der jedes Bild mit Metadaten wie Szene, Objekt, Kamerastandort und der Konfidenz der Klassifizierung versehen wurde. Auf der Grundlage dieser Metadaten einerseits und der algorithmisch aus den Einzelbildern extrahierten visuellen Merkmale (z. B. domi-

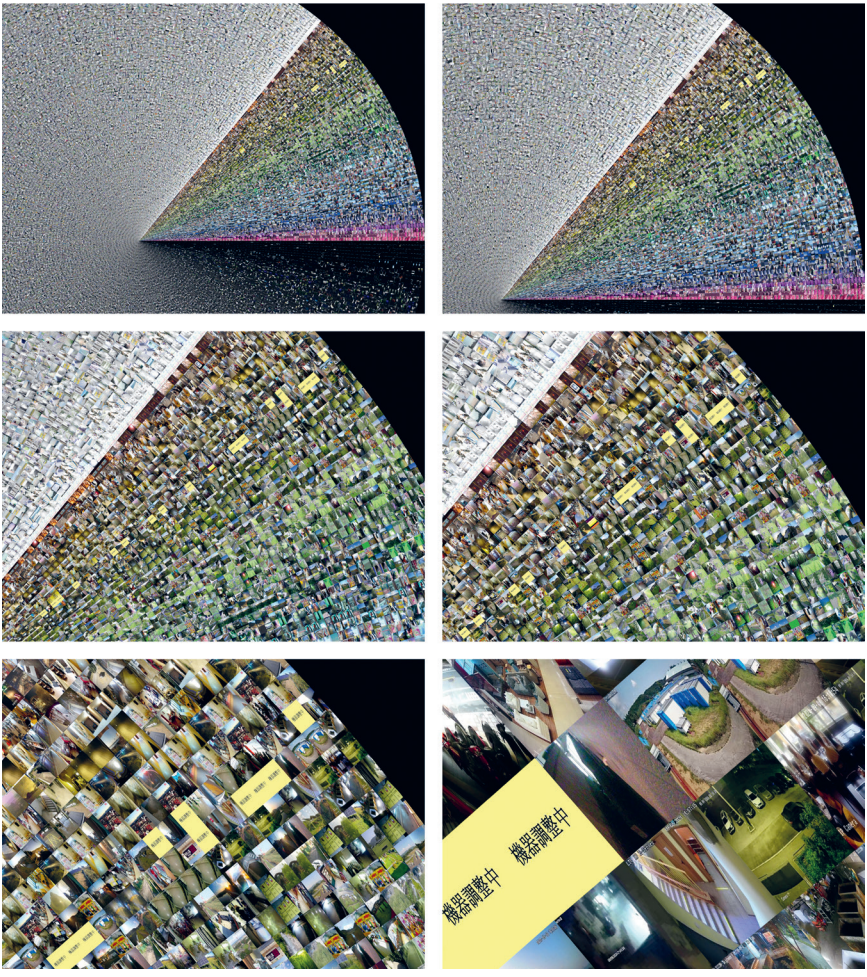
nante Farbe, Kontrast) andererseits können die Bilddaten sodann auf der Suche nach Ordnungsmustern variabel sortiert werden, was potenziell zu einer Vielzahl und Vielfalt an Metabildern führen kann.[6] Ausschlaggebend für die Erstellung des hier gezeigten Beispiels war das spezifische Interesse an Kamerabildern aus Asien in einer frühen Phase des Forschungsprozesses, insbesondere die Frage, inwiefern diese Bilder Praktiken der Überwachung zeigen und welche Einblicke sie in Machtverhältnisse, Zensur und soziale Kontrolle eröffnen. Zudem war die Intention ausschlaggebend, durch ein kreisförmiges, nach Farbton und Helligkeit sortiertes Metabild relevante Informationen und Bedeutungsstrukturen in Bezug auf die anfänglich fokussierten Forschungsfragen deutlich sichtbar machen zu können. Dabei wurde unter anderem vermutet, dass die bereits auf Netzkamerabildern aus anderen Kontinenten beobachtete visuelle Praxis der Verfremdung von Bildsegmenten durch Abdeckungen, Unschärfen oder Verpixelungen – etwa in Form von Balken, Weichzeichnungen oder groben Pixeln (vgl. Caviezel, 2015, S. 195 ff.) – besonders effizient identifiziert werden könnte, um einerseits deren konkrete Umsetzung in Asien zu erkunden und andererseits bestehende Forschungsfragen zu schärfen sowie neue zu generieren.

Die interaktive Bedienbarkeit des exemplarischen Metabilds wiederum unterstützt sowohl den Explorationsprozess nach relevanten Informationen und Bedeutungsstrukturen als auch die Entwicklung und Präzisierung von Forschungsfragen, indem sie einen fließenden Übergang von einem Gesamtüberblick über das visuelle Datenkorpus, wie in Abbildung 1 dargestellt, bis hin zu detaillierten Einblicken in spezifische Ausschnitte einzelner Bilddaten ermöglicht, wie in Abbildung 2 exemplarisch anhand von Screenshots einer Zoom-Bewegung illustriert. Auf diese Weise bietet das Beispielbild nicht nur eine umfassende Übersicht über alle 40 000 Bilder, sondern zeigt auch tiefere Strukturen, wie etwa eine Anomalie im Regenbogen-Farbverlauf: eine unterbrochene, gelb leuchtende Linie, die sich vom Mittelpunkt des Kreises bis zum Rand erstreckt. Zugleich ermöglicht es die gezielte Fokussierung auf einzelne Netzkamerabilder, was beim Hineinzoomen auch die besagte Anomalie näher erklärt: Die gelbe Linie resultiert aus identischen Bildern, auf denen schwarze chinesische Schriftzeichen auf gelbem Grund den Hinweis „Maschine wird gewartet“ zeigen.

Im Gegensatz zur Nutzung des Metabildes im Kontext des transdisziplinären Forschungsprojekts, wo es vor allem dazu dient, mit Blick auf das zentrale Forschungsinteresse an der Beobachtung der Welt via Netzkameras soziologisch relevante und/oder künstlerisch vielversprechende Strukturen in den 40 000 Netzkamerabildern zu identifizieren – sei es durch die Fokussierung auf die (Un-)Sichtbarkeit sozialer Situationen, Beziehungen und Räume, sei es durch den Vergleich mit Metabildern, die Bilder anderer Kontinente oder Zeitpunkte zeigen, um regionale Unterschiede und zeitliche Entwicklungen zu analysieren, oder sei es durch die Bewertung der

6 Diese Vielzahl und Vielfalt wird im Rahmen des Forschungsprojekts zusätzlich dadurch erweitert, dass entsprechende Metabilder auch für Netzkameras von anderen Kontinenten oder geografischen Standorten erstellt werden können.

Abbildung 2 Exemplarisches Hineinzoomen in das Metabild



Quelle: Eigene Visualisierung [Grabner und Hoggenmüller].

Potenziale und Grenzen des Metabildes als analytisches Werkzeug einschließlich technischer Fehler und ethischer Herausforderungen –, möchten wir im vorliegenden Beitrag anhand des exemplarischen Metabildes verdeutlichen, dass das, was wir als Metabild wahrnehmen, auf komplexen statistischen Berechnungen, algorithmischen Verfahren, technischen Bedingungen und einer Vielzahl von Entscheidungen basiert, die grundsätzlich kontingent und veränderbar sind. Und speziell an dieser Stelle: Es geht uns darum, aufzuzeigen, dass Forschende im Herstellungsprozess eines

Metabildes an verschiedenen Stellen Wahlmöglichkeiten haben und damit Entscheidungszwängen unterliegen, die die Datenvisualisierung entscheidend beeinflussen.

In unserem Beispiel betrifft diese Kontingenz etwa die Auswahl der Bilddaten (hier eine zufällige Stichprobe von 40 000 in Asien an einem Tag produzierten Bildern), die Wahl der Analysemodelle (eine Kombination aus drei Technologien), die Festlegung der Sortierkriterien (radiale Anordnung nach Farbton und Helligkeit), die Form der visuellen Darstellung (kreisförmig und zweidimensional), die Einbindung von Metadaten (Kamerastandort) sowie die Interaktivität und Benutzer*innenführung (Schwenk- und Zoomfunktion). Es ist leicht vorstellbar, dass eine andere Auswahl der Bilddaten (z. B. eine Variation der Zeitspanne oder eine gezielte inhaltliche Fokussierung auf Naturbilder oder Porträts), eine abweichende Wahl der Analysemodelle (z. B. Objektdetektion oder Szenenklassifikation), eine alternative Festlegung der Sortierkriterien (z. B. allein nach Bildinhalt oder allein nach Metadaten), eine veränderte visuelle Darstellung (z. B. als Cluster- oder Streudiagramm), der Einbezug anderer Metadaten (wie Zeitstempel oder Aufnahmewinkel) sowie variierende Interaktionsmöglichkeiten und Benutzer*innenführungen (z. B. Drehfunktion oder Ebenenverwaltung) zu grundlegend anderen Metabildern geführt hätten. Dadurch wäre maßgeblich beeinflusst worden, welche Muster sichtbar und welche Bedeutungsebenen hervorgehoben werden und wie flexibel die Bilddaten exploriert werden können.

All diese vorausliegenden Wahlmöglichkeiten und Entscheidungen bleiben in der Regel in Black Boxes verborgen – zumindest werden sie bei der Beschreibung von Forschungsergebnissen nur selten thematisiert und noch seltener kritisch hinterfragt –, sodass Metabilder trotz ihres kontingenten und entscheidungsabhängigen Herstellungsprozesses meist als evidente und objektive Darstellungen präsentiert werden. Dies ist umso bemerkenswerter, wenn man bedenkt, dass Metabilder nicht einfach bestehende Zusammenhänge abbilden, sondern Bedeutungsstrukturen wie etwa wiederkehrende Muster, verwandte Cluster oder isolierte Ausreißer in großen digitalen visuellen Datenbeständen überhaupt erst hervorbringen. Um diese Hervorbringung in ihrer Eigenheit genauer zu verstehen und darauf aufbauend den epistemischen Gehalt von Metabildern angemessen bewerten zu können, ist ein tieferes Verständnis der zugrunde liegenden Prozesse, Algorithmen, Standards und Praktiken sowie ihres Zusammenspiels notwendig. Anders formuliert: Wenn man Metabilder als Forschungswerkzeuge zur Erkenntnisgewinnung nutzen möchte, ist es notwendig, die Black Box der Metabilder zu öffnen, um besser zu verstehen, was Metabilder zeigen, was sie verbergen und was sie glauben machen. Im folgenden Abschnitt werden wir dies speziell mit Blick auf einen Aspekt tun, der in der Forschungspraxis oft schwer fassbar ist, weil er sich der Wahrnehmung in ganz besonderer Weise entzieht: die algorithmische Bedingtheit von Metabildern.[7]

7 Ergänzend dazu wird derzeit eine ethnografische Untersuchung vorbereitet, mit der im Kontext des Projekts *Watching the World* das kommunikative Handeln und dessen soziotechnische Bedingtheit bei der Herstellung, Verwendung und Interpretation von Metabildern in den Fokus genommen werden soll.

3 Einblicke in die Black Box – die Rolle von Algorithmen

Bei der Beschreibung der algorithmischen Bedingtheit von Metabildern liegt unser Schwerpunkt weniger auf den mathematischen Grundlagen der Algorithmen. Vielmehr möchten wir einen Einblick in die Pipeline[8] der Datenanalyse geben und die algorithmische Analyse visueller Ähnlichkeiten näher beleuchten: Nach welchen Logiken der Automation kommen Metabilder zustande? Welche epistemologischen Implikationen ergeben sich daraus? Oder konkret am Beispiel des im Abschnitt zuvor gezeigten Metabildes gefragt: Was bedeutet visuelle Ähnlichkeit in der computergestützten und algorithmusbasierten Datenvisualisierung der 40 000 Netzkamerabilder? Beim Erkunden des Metabildes erkennt man beispielsweise Cluster von Kamerabildern mit geringem Kontrast, während kontrastreiche Kamerabilder an anderer Stelle zu finden sind. Doch bei genauerer Betrachtung drängt sich die Frage auf, was die räumlich nahe beieinander liegenden Bilder tatsächlich gemeinsam haben, und umgekehrt, warum ähnlich erscheinende Bilder mitunter weit voneinander entfernt angeordnet sind. Diese Fragen leiten unsere Ausführungen zur Rolle von Algorithmen bei der Herstellung von Metabildern an, die wir anhand des mathematisch-statistischen Konzepts des Merkmalsraums (Abschnitt 3.1) und des Verfahrens der Dimensionsreduktion (Abschnitt 3.2) erläutern.

3.1 Merkmalsraum (Feature Space)

Zur Herstellung von Metabildern müssen Algorithmen zunächst sogenannte Features aus dem zugrunde liegenden Datensatz extrahieren. Im Kontext der Bildverarbeitung sind Features messbare Merkmale, die wesentliche Eigenschaften der Bildobjekte erfassen. Ziel dieses Prozesses ist es, die wichtigsten Informationen abstrahiert und komprimiert darzustellen, um die Datenmenge zu reduzieren und die Komplexität zu verringern. Dies schafft eine handhabbare Grundlage für die weitere Analyse der Daten (vgl. Bishop, 2006).

Für die Merkmalsextraktion (Feature Extraction) gibt es verschiedene Ansätze (vgl. als Überblick Balan P & Sunny, 2018). Diese können erstens auf handcodierte, das heißt vom Menschen definierte Features (klassische Bildfeatures) fokussieren, wobei grundsätzlich zwischen Features auf niedriger und auf hoher Abstraktionsebene unterschieden wird, auch bekannt als Low-Level-Features und High-Level-Features. Low-Level-Features umfassen grundlegende visuelle Eigenschaften, die direkte, pixelnahe Informationen wie Farben oder Kanten repräsentieren. Kanten bezeichnen dabei Stellen im Bild, an denen sich die Helligkeit oder Farbe stark ändert – sie markieren oft die Grenzen von Objekten (Bildelementen). Bei High-Level-Features

8 Der Begriff Pipeline bezeichnet strukturierte Arbeitsabläufe, die Daten durch eine festgelegte Abfolge von Schritten verarbeiten. Diese Schritte, die oft automatisiert ablaufen, zielen darauf ab, Daten zu sammeln, zu bereinigen, zu transformieren und schließlich zu analysieren. Das Ergebnis eines jeden Schritts dient dabei als Ausgangspunkt für den nächsten, wodurch ein durchgängiger und effizienter Prozess entsteht.

handelt es sich hingegen um komplexere Strukturen, die durch fortgeschrittene Bildverarbeitungsmethoden wie Segmentierung oder die Kombination mehrerer Merkmale aus den Bildpixeln gewonnen werden, beispielsweise Objekterkennung oder semantische Features, die Bedeutungen aus dem Bildinhalt ableiten (etwa die Erkennung von Emotionen in Gesichtern).

Zweitens können Features durch Convolutional Neural Networks (CNN), eine Klasse von Deep-Learning-Netzwerken, aus den Bilddaten extrahiert werden, wobei die Features in Form von Embeddings – komprimierten numerischen Repräsentationen der Bilddaten – ausgegeben werden. CNNs bestehen aus sogenannten Convolutional Layers, speziellen Schichten, die lokale Bildbereiche mithilfe von Filtern (auch Kernels genannt) analysieren und charakteristische Muster wie Kanten oder Texturen in numerische Werte umwandeln. Diese Schichten sind hierarchisch angeordnet, sodass CNNs während des Trainingsprozesses zunehmend komplexere Features lernen, die oft über klassische, handcodierte Bildfeatures hinausgehen. Im Unterschied zu Letzteren erfolgt die Feature Extraction bei CNNs automatisiert und datengetrieben, also ohne, dass die Features explizit definiert werden müssen. Vielmehr lernen die Netzwerke während des Trainingsprozesses, welche Features für eine bestimmte Aufgabe am nützlichsten sind. Gleichzeitig bleibt die interne Struktur des Modells dabei weitgehend unzugänglich: Die Prozesse, durch die das Netzwerk Features aus den Bilddaten extrahiert und in Embeddings überführt, sind aufgrund ihrer datengetriebenen Natur nicht direkt steuerbar. Ferner hängen die resultierenden Embeddings sowohl von den Trainingsdaten als auch der Netzwerkarchitektur ab und ermöglichen daher lediglich die Reproduktion der während des Trainings gelernten Repräsentationen, ohne dass die internen Repräsentationen vollständig kontrolliert werden können.

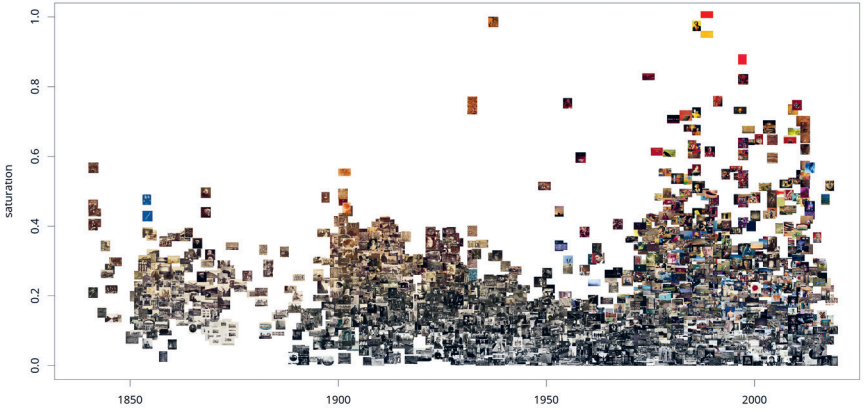
Drittens können Features auch aus vorhandenen Metadaten extrahiert werden, also aus Informationen, die nicht unmittelbar aus den Bildpixeln stammen, sondern die Bilddaten kontextuell ergänzen. Im Zusammenhang mit visuellen Daten aus Open-Data-Quellen wie dem bereits erwähnten Netzwerkkamera-Projekt oder Museumssammlungen können Metadaten verschiedene Informationen bereitstellen: im Fall der Netzwerkkameras Informationen wie den Kamerastandort, die Aufnahmezeit und die Wetterbedingungen, im Fall der Museumssammlungen den*die Künstler*in, das Entstehungsjahr und die Größe eines Kunstwerks. Diese und weitere Metadaten können in beiden Fällen als Features verwendet werden, um Ähnlichkeiten zwischen Netzwerkkamerabildern bzw. zwischen Kunstwerken zu erkennen, Muster und Trends im zeitlichen Verlauf der kamerabasierten Beobachtung der Welt bzw. in der Kunstgeschichte zu analysieren und Netzwerkkamerabilder bzw. Kunstwerke nach bestimmten Kriterien zu kategorisieren oder zu filtern.

Die Wahl zwischen Low-Level-Features und High-Level-Features, datengetriebenen Embeddings aus CNNs und metadatenbasierten Features hängt in der kultur-, geistes- und sozialwissenschaftlichen Forschungspraxis stark vom spezifischen

Forschungskontext, von den verfügbaren Daten und von den Zielen der Analyse ab. Grundsätzlich ist allen drei Ansätzen jedoch gemein, dass jedes Bildobjekt in einen n-dimensionalen Merkmalsraum, den sogenannten Feature Space, projiziert wird, wobei n die Anzahl der extrahierten Features oder die Dimension der Embeddings zur Beschreibung eines Bildobjekts angibt.

Was hier abstrakt klingt – die Projektion von Bildobjekten in einen Feature Space –, möchten wir an einem Beispiel veranschaulichen. Wir nutzen dafür die vom Museum of Modern Art (MoMA) als Open Data bereitgestellten Sammlungsdaten (vgl. MoMA o. J.). Dieser Datensatz ist für unsere Illustrationszwecke besonders geeignet, da er neben den Digitalisaten der Kunstwerke systematisch gepflegte und kuratierte Metadaten von hoher Qualität und Konsistenz bietet. Konkret enthält er eine Vielzahl textbasierter Features (wie Titel, Künstler*in, Abteilung) und numerischer Features (wie Entstehungs- und Erwerbsdatum, Höhe, Breite). Dank dieser reichhaltigen und vielfältigen Features können die Kunstwerke präzise und differenziert in einen mehrdimensionalen Feature Space projiziert werden. Um das Verfahren der Projektion verständlich darzustellen, beschränken wir uns bewusst auf die in der Sammlung enthaltenen Fotografien als Bildobjekte und wählen nur zwei Features: das Entstehungsjahr, das aus den Metadaten entnommen wird, und die Durchschnittssättigung, die als visuelles Feature direkt aus den Bilddaten abgeleitet wird und den mittleren Sättigungswert über alle Pixel eines Bildes beschreibt. Diese beiden Features ermöglichen es, die Bildobjekte in einem zweidimensionalen XY-Diagramm abzubilden (Abb. 3).

Abbildung 3 Metabild nach Entstehungsjahr und Durchschnittssättigung, Fotografien des MoMA



Quelle: Eigene Visualisierung [mit R erstelltes Metabild, Klinke].

Das Metabild in Abbildung 3 zeigt die Verteilung der Fotografien des MoMA entlang der beiden Achsen Entstehungsjahr und Durchschnittssättigung. Auffällig ist dabei, dass bereits zwei Features genügen, um eine plausible, aussagekräftige Anordnung hinsichtlich der visuellen Ähnlichkeit der Bildobjekte zu erzielen. So lässt sich in dem Metabild beispielsweise das Aufkommen der Sepia-Fotografie in der Mitte des 19. Jahrhunderts erkennen. Erst später wird die Schwarz-Weiß-Fotografie zunehmend dominant, insbesondere das Gelatine-Silber-Verfahren, das am unteren Ende der Sättigungsskala angesiedelt ist. In der zweiten Hälfte des 20. Jahrhunderts kommen schließlich Farbabzüge hinzu, die sich über den gesamten Sättigungsbereich verteilen. Darüber hinaus offenbart die Visualisierung eine Sammlungslücke zwischen etwa 1870 und 1890, die möglicherweise überhaupt erst durch diese quantitative Analyse sichtbar wird und deren Ursachen weiterführend mit qualitativen Methoden eruiert werden könnten. Allgemein ist festzuhalten: Obwohl die erkennbaren Cluster ausschließlich auf den beiden von uns exemplarisch ausgewählten Features basieren, weisen sie an unterschiedlichen Stellen semantische Zusammenhänge auf, die zwar nicht explizit in den Features codiert sind, aber mit ihnen korrelieren. Eine ausführlichere Diskussion hierzu folgt im nächsten Abschnitt (3.2).

Jenseits unseres Beispiels ist ein n-dimensionaler Feature Space ein abstraktes mathematisch-statistisches Konzept, das dazu dient, die Eigenschaften von Datenpunkten in einem definierten Raum zu beschreiben. Jede Dimension dieses Raums steht für ein spezifisches Feature der Daten. Analysiert man beispielsweise Bilder nicht (wie in unserem Beispiel) anhand von zwei, sondern anhand von drei Features (etwa Helligkeit, Kontrast und Anzahl bestimmter Kanten), so lässt sich jedes Bildobjekt als Datenpunkt in einem dreidimensionalen Raum darstellen, wobei die x-, y- und z-Achse jeweils eines dieser Features repräsentiert. Werden für jedes Bildobjekt wiederum mehrere solcher Features ausgewählt, so kann es als Datenpunkt in einem n-dimensionalen Raum verstanden werden. Wenn zum Beispiel 4096 Features aus einem Bild extrahiert werden, wie es bei vortrainierten neuronalen Netzen wie VGG16 der Fall ist, wird das Bildobjekt in einem 4096-dimensionalen Raum positioniert. Dies ist zwar nichts, was man sich visuell vorstellen muss (oder kann), es bedeutet aber: Je mehr Features verwendet werden, desto höher ist die mathematische Präzision bei der Beschreibung eines Bildobjekts und der Differenzierung zu anderen Bildobjekten. Diese Präzision ist jedoch unmittelbar an den Korpus des vortrainierten Modells gebunden, das heißt, die Genauigkeit der Beschreibung bleibt auf jene Features und Kategorien beschränkt, die das Modell während des Trainings kennengelernt hat und als Embeddings im hochdimensionalen Raum darstellt.[9]

Entscheidend für unsere Beschreibung der algorithmischen Bedingtheit von Metabildern ist hierbei, dass die Positionierung der Datenpunkte im Feature

9 Wenn beispielsweise Schlagwörter vergeben werden, können nur die Schlagwörter aus dem Korpus des Netzes genutzt werden. Ebenso wird das Modell Schwierigkeiten haben, Kategorien zu erkennen, die nicht im Trainingskorpus enthalten sind.

Space weiterführende Berechnungen ermöglicht. Beispielsweise können damit die Abstände zwischen Objekten berechnet werden, die eine zentrale Rolle bei der Bestimmung ihrer Ähnlichkeit spielen. Die Datenpunkte werden dabei als Vektoren im n-dimensionalen Raum dargestellt, deren Beziehungen durch verschiedene mathematische Operationen wie Abstandsmessungen oder Winkelberechnungen analysiert werden können. Der euklidische Abstand misst die direkte Distanz zwischen zwei Punkten und eignet sich besonders, wenn die physische Nähe der Datenpunkte von Bedeutung ist. Alternativ kann die Kosinus-Ähnlichkeit verwendet werden, die den Winkel zwischen den Vektoren berücksichtigt, was sinnvoll ist, wenn die Richtung der Vektoren wichtiger ist als ihre absolute Länge. Die so berechneten Abstandswerte geben an, wie stark sich zwei Objekte in Bezug auf ihre Features ähneln oder unterscheiden. Ein kleiner Abstand deutet darauf hin, dass die Objekte ähnliche Features haben, während ein großer Abstand anzeigt, dass die Objekte stark voneinander abweichen. Solche Abstandsmessungen sind zentral für viele algorithmusbasierte Anwendungen, darunter Klassifizierungen (vgl. etwa im Bereich Public Health Rösch, 2022), Clusterings (vgl. z. B. in der Literarkritik Tschuggnall et al., 2016) und Empfehlungssystemen (vgl. u. a. aus soziologischer Perspektive Unternährer, 2024). Im Zusammenhang mit Metabildern helfen sie wiederum, Beziehungen visuell darzustellen und dadurch Muster sowie Strukturen in großen visuellen Datenbeständen zu identifizieren.

3.2 Dimensionsreduktion (Dimension Reduction)

Wie im vorherigen Abschnitt dargelegt, eröffnet eine größere Anzahl von Features, die zur Beschreibung von Datenobjekten verwendet werden, potenziell vielfältigere statistische Möglichkeiten auf der Ebene der Daten. Doch auf der Ebene der Visualisierung stoßen wir schnell an Grenzen: Menschen sind daran gewöhnt, nur eine bestimmte Anzahl von Dimensionen wahrzunehmen. Die physische Welt, die wir erleben, umfasst im Wesentlichen drei Dimensionen, die sich in einem dreidimensionalen Raum gut visualisieren lassen, während Zeitlichkeit zusätzlich durch Animationen dargestellt werden kann.[10] Bei höherdimensionalen Datensätzen hingegen wird es zunehmend schwieriger, sie intuitiv zu erfassen oder visuell darzustellen.

Eine Möglichkeit, mit dieser Herausforderung umzugehen, sind mathematisch-statistische Verfahren der Dimensionsreduktion. Diese Verfahren zielen darauf ab, hochdimensionale Daten, also Daten, bei denen jedes einzelne Datenobjekt durch eine sehr große Anzahl an Features beschrieben wird, in einen niedrigdimensionalen Raum zu projizieren, wobei die wesentlichen Eigenschaften der Daten, insbesondere ihre Strukturen und Muster, erhalten bleiben. Dabei entstehen neue, repräsentative Dimensionen, die häufig als Kombinationen oder Transformationen der ursprüng-

10 Das Potenzial solcher Visualisierungen hat Hans Rosling eindrucksvoll gezeigt; vgl. etwa seine TED-Talks (https://www.ted.com/playlists/474/the_best_hans_rosling_talks_yo).

lichen Features gebildet werden. Dies verbessert nicht nur die Interpretierbarkeit, sondern auch die Effizienz und Anwendbarkeit der Daten für maschinelle Lernmodelle. Eine Alternative wäre, nur die für die Datenanalyse relevantesten Features beizubehalten und irrelevante oder weniger informative Features zu eliminieren, mithin eine bestimmte Untermenge der ursprünglichen Dimensionen auszuwählen, ein Prozess, der auch als Merkmalsauswahl (Feature Selection) bezeichnet wird.

Die Dimensionsreduktion ist prinzipiell vergleichbar mit der Projektion des Schattens eines dreidimensionalen Gegenstands auf eine flache Ebene: Wird ein dreidimensionaler Gegenstand beleuchtet, entsteht eine zweidimensionale Projektion, die eine vereinfachte Darstellung des Gegenstands ist, da in ihr die Tiefeninformation fehlt. Dabei können jedoch auch andere relevante Eigenschaften des Gegenstands verloren gehen. Um einen solchen Informationsverlust zu minimieren, wurden verschiedene Algorithmen entwickelt, die die Dimensionalität reduzieren, indem sie einen neuen, kompakteren Feature Space schaffen, der einerseits eine visuelle Darstellung ermöglicht, die von Menschen interpretiert werden kann, und andererseits die anderen Dimensionen so weit wie möglich bewahrt. Oder anders ausgedrückt: Die 4096 visuellen Features, die beispielsweise durch VGG16 extrahiert werden, werden in einen niedrigdimensionalen (üblicherweise zweidimensionalen) Raum projiziert, während bestimmte relevante Informationen erhalten werden. Das Ergebnis dieser Transformation ist eine XY-Position des Bildobjekts im neuen Feature Space.

Es gibt eine Reihe von Algorithmen zur Dimensionsreduktion, die sich unter anderem in ihren mathematischen Grundlagen, den Annahmen über die Datenstruktur und in der Art und Weise, wie sie mit Daten umgehen, unterscheiden. Zu den bekanntesten gehören die Principal Component Analysis (PCA), das t-Distributed Stochastic Neighbor Embedding (t-SNE) und die Uniform Manifold Approximation and Projection (UMAP), wobei alle drei jeweils einen eigenen Ansatz bieten, um die hochdimensionalen Daten in eine Form zu überführen, die sowohl der menschlichen Wahrnehmung zugänglich ist als auch von maschinellen Lernprozessen effizienter verarbeitet werden kann.[11] Der Algorithmus t-SNE hat sich in der kultur-, geistes- und sozialwissenschaftlichen Forschung als besonders nützliches Werkzeug für die explorative Datenanalyse und die Visualisierung komplexer Datensätze erwiesen, weshalb wir uns im Folgenden auf diesen Algorithmus konzentrieren.

Der Dimensionsreduktionsalgorithmus t-SNE basiert auf dem Stochastic Neighbor Embedding (SNE), das 2002 von Geoffrey Hinton und Sam Roweis

11 Dabei hat jeder dieser Ansätze seine Vor- und Nachteile, insbesondere in Bezug auf die Verarbeitungsgeschwindigkeit und die Art des Informationsverlusts. PCA erfasst nur lineare Zusammenhänge, ist jedoch der schnellste Ansatz, da er effizient berechnet wird und sich gut für große Datensätze skalieren lässt. UMAP bietet eine gute Balance zwischen Geschwindigkeit und der Erhaltung lokaler Strukturen, wobei es globale Zusammenhänge besser bewahrt als t-SNE. t-SNE wiederum ist langsamer, liefert jedoch eine sehr gute Visualisierung der lokalen Datenstruktur, indem es Cluster und Nachbarschaftsbeziehungen bewahrt, jedoch oft auf Kosten der globalen Struktur (vgl. für eine kritische Perspektive auf PCA Shen et al., 2012; auf UMAP Damrich & Hamprecht, 2021; auf t-SNE Wattenberg et al., 2016).

entwickelt wurde. Ziel von SNE war es, die lokalen Nachbarschaftsbeziehungen optimal zu erhalten, indem ähnliche Datenpunkte im hochdimensionalen Raum auch im niedrigdimensionalen Raum nahe beieinander platziert werden (vgl. Hinton & Roweis, 2003). 2008 wurde das SNE-Verfahren dann von Laurens van der Maaten und Geoffrey Hinton zu t-SNE weiterentwickelt, wobei sie im niedrigdimensionalen Raum eine t-Verteilung anstelle der Gaußschen Verteilung einführten. Da die t-Verteilung extreme Distanzen zwischen Datenpunkten besser erfasst, mindert diese Modifikation das Überfüllungsproblem (crowding problem) und verbessert die Visualisierung von Clustern in großen Datensätzen (vgl. van der Maaten & Hinton, 2008).

Um die Nachbarschaftsbeziehungen zwischen den Datenpunkten weitgehend zu erhalten, wandelt t-SNE die Ähnlichkeiten im hochdimensionalen Raum (mithilfe einer Gaußschen Verteilung) in Wahrscheinlichkeiten um und versucht, diese Wahrscheinlichkeiten im niedrigdimensionalen Raum (mithilfe einer t-Verteilung) möglichst genau nachzubilden. Dabei werden die Datenpunkte zunächst zufällig im niedrigdimensionalen Raum platziert. Anschließend passt der Algorithmus die Position der Datenpunkte schrittweise an, um die Diskrepanz zwischen den Wahrscheinlichkeiten im hoch- und im niedrigdimensionalen Raum zu minimieren. Dieser Prozess wird so lange fortgesetzt, bis die bestmögliche Übereinstimmung zwischen den Wahrscheinlichkeiten erreicht ist.

Generell liefert t-SNE dabei kein festes, deterministisches Ergebnis. Stattdessen variiert die genaue Anordnung der Datenpunkte bei jedem Durchlauf, da der Algorithmus mit zufällig gewählten Startpositionen beginnt und durch Zufallsfaktoren im Optimierungsprozess beeinflusst wird.[12] Dies bedeutet, dass der Algorithmus verschiedene Versionen der hochdimensionalen Datenpunkte in der niedrigdimensionalen Darstellung erzeugt, während die übergeordneten Muster in der Regel erhalten bleiben. Kurz gesagt: Wiederholte Ausführungen des Algorithmus führen zu unterschiedlichen visuellen Darstellungen.

Dies lässt sich erneut anhand der MoMA-Daten veranschaulichen: Zu Demonstrationszwecken berücksichtigen wir diesmal nur die ersten 1000 Bildobjekte des Datensatzes und nutzen ausschließlich Features aus den Metadaten, verzichten also auf aus den Bildern abgeleitete Features. Mithilfe der Programmiersprache und statistischen Umgebung R werden daraufhin aus den 29 Spalten der MoMA-Daten, die die verschiedenen Bildobjekte beschreiben, elf numerische Features[13] ausgewählt und in numerische Werte umgewandelt. Anschließend reduzieren wir diese Daten

12 Zumindest theoretisch könnte dieses Problem durch die Verwendung eines fixierten Startwerts (Random Seed) gelöst werden. In der Praxis nutzen jedoch die meisten Implementierungen Parallelisierungen auf der Central Processing Unit (CPU) oder Graphics Processing Unit (GPU), wodurch Berechnungen in variierender Reihenfolge und Geschwindigkeit ablaufen. Dies führt dazu, dass die exakte Reproduzierbarkeit aufgrund der parallel ablaufenden Prozesse oft nicht gewährleistet ist.

13 Die elf numerischen Features sind folgende: ObjectID, BeginDate, EndDate, DateAcquired, Depth, Diameter, Height, Weight, Width, Duration und AspectRatio.

mit dem Paket Rtsne, eine R-Implementierung der t-SNE-Methode, auf zwei Dimensionen, indem die gewünschte Anzahl an Dimensionen als Parameter übergeben wird. Das Ergebnis ist eine zweidimensionale Matrix mit X- und Y-Koordinaten, auf deren Grundlage die zugehörigen Bilder in einem Metabild dargestellt werden können (siehe Abb. 4). Die X- und Y-Koordinaten spiegeln jedoch keine Achsenwerte wider, wie man es von einem herkömmlichen Diagramm erwarten würde. Stattdessen transformiert der Prozess die Eigenschaften eines n-dimensionalen Raums in eine Darstellung, in der sich die Position eines Datenpunkts weniger aus einer absoluten Maßstabsskala als vielmehr aus den relativen Abständen und Ähnlichkeiten zu anderen Datenpunkten ergibt.

Abbildung 4 Anwendung von t-SNE auf die numerischen Features aus den Metadaten der ersten 1000 Objekte des MoMA-Datensatzes

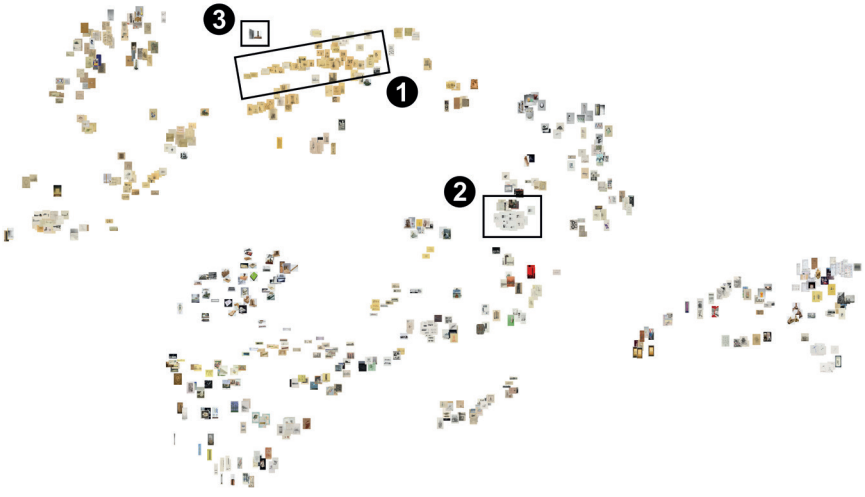
Quelle: Eigene Visualisierung [mit R erstelltes Metabild, Umrahmungen und Ziffern hinzugefügt, Hoggenmüller und Klinke].

Das Metabild zeigt strukturiert verteilte Miniaturbilder, die unterschiedlich viel Freiraum zueinander aufweisen. Teilweise gruppieren sich die Miniaturbilder nach Farben, wie zum Beispiel die Schwarz-Weiß-Zeichnungen des Central Parks von Bernard Tschumi, die oval auf der linken Seite angeordnet sind (siehe Umrahmung 2 auf der linken Seite des Metabildes), oder die bräunlichen Holzkohlezeichnungen auf Transparentpapier von Louis I. Kahn für eine Gebetshalle in Dhaka, die oben

rechts in einem leichten Bogen zu sehen sind (siehe Umrahmung 1 auf der rechten Seite des Metabildes). Diese Gruppierung ist vor allem deshalb bemerkenswert, weil Bildfeatures wie Farbigkeit nicht explizit berücksichtigt wurden. Vielmehr gilt: Die Platzierung der Miniaturbilder im reduzierten Feature Space lässt sich dadurch erklären, dass die Werke in einem gemeinsamen Kontext entstanden sind und daher visuelle Ähnlichkeiten aufweisen, die sich in ihrer räumlichen Nähe zueinander widerspiegeln. Es handelt sich also um eine Korrelation. Bei näherer Betrachtung ist die Platzierung der Fotografien von Modellen Ludwig Mies van der Rohes (siehe Umrahmung 3 auf der rechten Seite des Metabildes) in der Nähe der Werke von Louis I. Kahn im globalen Kontext der Metadaten jedoch schwer nachzuvollziehen, da die Werke in unterschiedlichen historischen und geografischen Kontexten entstanden sind.

Bei wiederholter Anwendung des t-SNE-Algorithmus entsteht ein davon abweichendes Metabild (siehe Abb. 5): t-SNE liefert aufgrund des oben beschriebenen probabilistischen Ansatzes und der zufälligen Initialisierung bei gleichem Input nicht dieselbe Ergebnismatrix. Obwohl das grundlegende Muster ähnlich bleibt, variiert bei der neuen Ausführung (und auch bei jeder weiteren Wiederholung) die Anordnung der Miniaturbilder.

Abbildung 5 Wiederholte Anwendung von t-SNE auf dieselben Features



Quelle: Eigene Visualisierung [mit R erstelltes Metabild, Umrahmungen und Ziffern hinzugefügt, Hoggenmüller und Klinke].

Diese Variationen bedeuten jedoch nicht, dass die Zusammenhänge völlig zufällig sind. Auch im zweiten Metabild finden sich die Werke von Kahn und Tschumi in Gruppen, diesmal allerdings zentraler platziert (siehe die Umrahmungen 1 und 2), und die Fotografien von Mies van der Rohes Modellen sind wieder in relativer Nähe zum Tschumi-Cluster angeordnet, wenn auch an anderer Stelle, nämlich links darüber (siehe Umrahmung 3).

Dieser kurze Vergleich der beiden Metabilder verdeutlicht, dass Visualisierungsmethoden wie t-SNE zwar hilfreich sind, um Beziehungen innerhalb großer visueller Datenbestände zu erkunden, ihre Ergebnisse in Form von Metabildern jedoch eine sorgfältige Interpretation und ein tieferes Verständnis der zugrunde liegenden Algorithmen erfordern. Außerdem wird auch an diesem Beispiel sichtbar, dass Forschende während des gesamten Analyseprozesses an verschiedenen Stellen Wahlmöglichkeiten haben und Entscheidungen treffen müssen, die das visuelle Ergebnis beeinflussen. In diesem Fall war es insbesondere die Auswahl der Objekte (die ersten 1000), der Features (ausschließlich numerische Features aus den Metadaten) sowie des Algorithmus zur Dimensionsreduktion (t-SNE). Mit anderen Worten: Es ist wichtig, zu beachten, dass t-SNE – wie alle Dimensionsreduktionstechniken – bestimmte Interpretationsherausforderungen mit sich bringt und die Ergebnisse immer im Kontext der Daten und unter Berücksichtigung der Parameterwahl sowie der Eigenheiten des spezifischen algorithmischen Verfahrens interpretiert werden sollten.

4 Metabilder und ihre kritische Nutzung – interdisziplinäre Herausforderungen und Zusammenarbeit

Unsere Fragen zum näheren Verständnis von Metabildern sind mit den bisherigen Ausführungen keineswegs abschließend geklärt. Vielmehr eröffnen sich daran anschließend neue Perspektiven für eine kritische Nutzung von Metabildern als Forschungswerkzeug in den Kultur-, Geistes- und Sozialwissenschaften, die mit einer Reihe von Herausforderungen einhergeht. Um diesen Herausforderungen produktiv zu begegnen, bedarf es vertiefter interdisziplinärer Zusammenarbeit, die über unsere kooperative Herangehensweise aus Visueller Soziologie und Digitaler Bildwissenschaft hinausreicht. Drei dieser Herausforderungen, die wir als besonders zentral erachten, stehen im Fokus unserer abschließenden Überlegungen.

Die erste Herausforderung besteht in der Mehrdimensionalität von Metabildern: Metabilder vereinen mehrere Informationsebenen, die von ihrer interaktiven, visuellen Darstellungsform bis hin zu den ihnen zugrunde liegenden visuellen Daten und Metadaten reichen. Diese verschiedenen Ebenen erfordern unterschiedliches fachspezifisches Wissen sowie unterschiedliche analytische Ansätze; erst deren Kombination ermöglicht ein umfassenderes Verständnis der Metabilder und ihrer

Komplexität. So bedarf es unter anderem fundierter Fachkenntnisse in der Analyse von Einzelbildern, insbesondere im Verstehen ihrer symbolischen Bedeutungen und historischen Kontexte (z. B. mittels Ikonografie, Ikonologie, visueller Semiotik), um spezifische Inhalte und visuelle Merkmale zu identifizieren und visuelle Narrative zu entschlüsseln. Notwendig sind aber auch methodische Kompetenzen im Bildvergleich, um visuelle Muster, Stile und Kompositionen aus komparativer Perspektive zu analysieren. Ebenso essenziell ist ein technisches Wissen, beispielsweise in der Bildverarbeitung (z. B. Computer Vision, CNNs), in der Programmierung (Sprachen wie Python oder R) sowie im Umgang mit Bibliotheken für maschinelles Lernen und Datenvisualisierung (etwa TensorFlow und Matplotlib), um die Prozesse der Mustererkennung und Anomalieentdeckung in großen visuellen Datensätzen zu verstehen, zu reflektieren und gegebenenfalls weiterzuentwickeln. Ferner bedarf es eines Wissens über konzeptionelle Rahmen – oder die Fähigkeit, solche Rahmen selbst zu entwickeln –, um die Qualität der den Metabildern zugrunde liegenden großen Datensätze zu bewerten. Dies betrifft die Metadaten, aber auch die visuellen Daten selbst, mit besonderem Augenmerk auf die Identifikation von Daten minderer Qualität sowie die Entwicklung von Strategien zur Verbesserung der Datenqualität (vgl. hierzu im Bereich der computergestützten Textanalyse Hurtado Bodell et al., 2022). Und nicht zuletzt ist profundes Wissen über kulturelle, soziale und historische Implikationen entscheidend, um die in Metabildern sichtbaren Muster und Strukturen tiefergehend zu verstehen und zu erklären – oder sie als heuristische Grundlage für weiterführende qualitative Forschungen fruchtbar zu machen.

Die zweite Herausforderung liegt in der stochastischen Natur der Algorithmen zur Bildverarbeitung und visuellen Mustererkennung. Wie wir gezeigt haben, führt diese dazu, dass bei wiederholter Anwendung unterschiedliche Metabilder desselben Datensatzes entstehen können. Der von uns beschriebene t-SNE-Algorithmus, dessen *S* für *stochastic* steht, verdeutlicht die inhärente Unbestimmtheit solcher Methoden und unterstreicht die Notwendigkeit, die scheinbare Evidenz von Metabildern kritisch zu hinterfragen, nicht zuletzt im Hinblick auf die wissenschaftlichen Kriterien der Reproduzierbarkeit und der Reliabilität der Ergebnisse. Die Fähigkeit dieser Algorithmen, bereits mit wenigen Features signifikante Korrelationen zu identifizieren und visuelle Cluster zu bilden, weist zudem auf die Gefahr hin, voreilige Schlüsse aus Metabildern zu ziehen. Entsprechend erfordert die Beurteilung der Epistemik von Metabildern sowie die kritische Auseinandersetzung mit auf Ähnlichkeit basierenden Systemen einerseits eine interdisziplinäre Perspektive, die nicht nur ein mathematisches Verständnis der Algorithmen einschließt, sondern auch technologische Entwicklungen in der Bildverarbeitung und die sozio-historische Kontextualisierung jener Evolutionen mit epistemologischen und ethischen Überlegungen verknüpft. Eine enge Zusammenarbeit zwischen Forschenden aus Bereichen wie Informatik, Statistik, Bildwissenschaft, Soziologie, Ethik und Wissenschaftsgeschichte ist entscheidend, um die epistemischen Chancen und Risiken dieser Technologien umfas-

send zu verstehen und Fehlinterpretationen zu vermeiden. Andererseits muss bereits die Forschungspraxis in den Kultur-, Geistes- und Sozialwissenschaften über die reine Beschreibung hinausgehen und technisches Experimentieren aktiv integrieren. Technisches Experimentieren bezieht sich in diesem Kontext auf das Ausprobieren und Anpassen von verschiedenen Mapping-Verfahren zur Datenvisualisierung (vgl. zum Prozess des Mappings in der Design-Based Research etwa Schranz, 2021; speziell in der Soziologie Marguin et al., 2024), um deren Auswirkungen auf die Darstellung und Interpretation von Daten besser zu verstehen. Durch das Testen unterschiedlicher Parameter und Techniken können die Stärken und Schwächen der jeweiligen Methoden evaluiert werden, was zu einer kritischeren Hinterfragung und einem differenzierteren Verständnis der visuellen Ergebnisse führt.

Die dritte Herausforderung ist die fortlaufende Weiterentwicklung von Algorithmen, die bei der Erstellung von Metabildern wesentlich sind. Neben den allgemeinen Zugangsproblemen, denen die Forschung zu Algorithmen häufig begegnet – insbesondere wenn die Algorithmen von privaten Unternehmen kontrolliert werden (vgl. Kitchin, 2016; Herms & Lehmann in diesem Sonderheft) –, besteht das Problem eines stetig wachsenden Bedarfs an technischen Kompetenzen. Ein interdisziplinärer Ansatz ist daher auch vor diesem Hintergrund unverzichtbar (vgl. zum Feld der Critical Data Studies jüngst Kitchin, 2025): Technische und mathematische Kompetenzen müssen systematisch mit den Fragestellungen aus den Kultur-, Geistes- und Sozialwissenschaften verknüpft werden, um die Arbeit mit Metabildern weiter zu optimieren und die Potenziale technischer Entwicklungen voll auszuschöpfen. Exemplarisch ist mit Blick auf die Visualisierung von Ähnlichkeiten denkbar, dass Metabilder nicht auf zwei- oder dreidimensionale Räume beschränkt bleiben müssen. Neue immersive Darstellungsformen – beispielsweise basierend auf Virtual Reality (VR), Augmented Reality (AR) oder Mixed Reality (MR) – eröffnen vielversprechende Wege, Zusammenhänge zwischen visuellen Daten auf innovative Weise buchstäblich erfahrbar zu machen: Mithilfe von Extended Reality (XR) könnten Forschende nochmals tiefer in große visuelle Datenbestände eintauchen und dadurch möglicherweise neue Erkenntnisse generieren (vgl. richtungsweisend mit Blick auf videobasierte qualitative Untersuchungen McIlvenny & Davidsen, 2017). Die Entwicklung solcher immersiven Forschungsumgebungen, die virtuelle und physische Welten miteinander verbinden, erfordert eine kontinuierliche und enge Zusammenarbeit aller Beteiligten. Dazu gehört insbesondere der Dialog zwischen Forschenden aus den Kultur-, Geistes- und Sozialwissenschaften, die spezifische Anforderungen an die Visualisierungen haben, und den Fachleuten aus Bereichen wie Computer Vision, maschinelles Lernen, Data Science, IT-Infrastruktur, Interaktionsdesign (Interaction Design) und nutzer*innenzentriertes Design (User-Centered Design). Gemeinsam müssen sie die technischen Möglichkeiten nicht nur weiterentwickeln, sondern auch sicherstellen, dass diese den Bedürfnissen und Zielsetzungen der kultur-, geistes- und sozialwissenschaftlichen

Forschung entsprechen. Nur ein solch interdisziplinärer Ansatz wird es ermöglichen, zukunftsfähige digitale Werkzeuge, Plattformen und Umgebungen für die Sammlung, Analyse und Visualisierung großer Mengen digitaler visueller Daten zu konzipieren und zu realisieren. Dies ist von entscheidender Bedeutung, da die Wahl und Gestaltung der Visualisierungsmethode – dies wollten wir mit unserem Artikel nicht zuletzt deutlich machen – maßgeblich beeinflusst, wie Wissen aus großen visuellen Datenbeständen extrahiert wird und wie Informationen präsentiert und interpretiert werden. Gleichzeitig kann erst im Zusammenspiel von mathematischem Verständnis, technischem Experimentieren, theoretischer Analyse und historischer, kultureller und sozialer Kontextualisierung der epistemische Wert von Metabildern präziser verstanden, reflektiert genutzt und vollständig gewürdigt werden – auch über die Kultur-, Geistes- und Sozialwissenschaften hinaus.

5 Literatur

Abdullahu, F., & Grabner, H. (2024). Commonly Interesting Images. *arXiv*. Preprint / Working Paper. https://doi.org/10.48550/arXiv.2409.16736

Balan P, S., & Sunny, L. E. (2018). Survey on Feature Extraction Techniques in Image Processing. *International Journal for Research in Applied Science and Engineering Technology*, *6*(3), 217–222. https://doi.org/10.22214/ijraset.2018.3035

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer Science+Business Media.

Caviezel, K. (2015). *The Encyclopedia of Kurt Caviezel*. Rorhof.

Caviezel, K., & Zürcher Hochschule für Angewandte Wissenschaften. o. J. Watching the World. *ZHAW*. https://webcamaze.engineering.zhaw.ch/ (letzter Zugriff am 30. Oktober 2024).

Damrich, S., & Hamprecht, F. A. (2021). On UMAP's True Loss Function. *arXiv*. Preprint / Working Paper. https://doi.org/10.48550/arXiv.2103.14608

Frischknecht, M. (2025). Through the Eyes of the Machine: Exploring Historical Photo Collections with Convolutional Neural Networks. *Schweizerische Zeitschrift für Soziologie, 51*(2), Special Issue hrsg. von S. W. Hoggenmüller, Big Visual Data als neue Form des Wissens: Potenziale, Herausforderungen und Transformationen.

Herms, K., & Lehmann J. (2025). Seeing Like a Field? *Schweizerische Zeitschrift für Soziologie, 51*(2), Special Issue hrsg. von S. W. Hoggenmüller, Big Visual Data als neue Form des Wissens: Potenziale, Herausforderungen und Transformationen.

Hinton, G. E., & Roweis, S. (2003). Stochastic Neighbor Embedding. In S. Becker, S. Thrun, & K. Obermayer (Hrsg.), *Advances in Neural Information Processing Systems 15: Proceedings of the 2002 Conference* (S. 857–864). MIT Press.

Hochman, N., & Schwartz, R. (2021). Visualizing Instagram: Tracing Cultural Visual Rhythms. *Proceedings of the International AAAI Conference on Web and Social Media*, *6*(4), 6–9. https://doi.org/10.1609/icwsm.v6i4.14361

Hoggenmüller, S. W. (2016). Die Welt im (Außen-)Blick: Überlegungen zu einer ästhetischen Re|Konstruktionsanalyse am Beispiel der Weltraumfotografie ‚Blue Marble'. *Zeitschrift für Qualitative Forschung, 17*(1–2), 11–40. https://doi.org/10.3224/zqf.v17i1-2.25541

Hoggenmüller, S. W. (2022). *Globalität sehen: Zur visuellen Konstruktion von „Welt"*. Campus. https://doi.org/10.12907/978-3-593-44235-8

Hoggenmüller, S. W. (2025). Mit den Händen denken: Ästhetisch-praktische Verfahren in der (sozial) wissenschaftlichen Bildanalyse. In M. Rosenkranz & N. T. Zahner (Hrsg.), *Plurale Verschränkungen: Zur Entdifferenzierung von Kunst, Politik, Wissenschaft und Wirtschaft* (S. 143–168). Springer VS. https://doi.org/10.1007/978-3-658-45684-9_8

Hoggenmüller, S. W., Caviezel, K., Abdullahu, F., & Grabner, H. (In Begutachtung). *Watching the World: Big Visual Data Research in the Dialogue between Art, Computer Vision, and Sociology*.

Hurtado Bodell, M., Magnusson, M., & Mützel, S. (2022). From Documents to Data: A Framework for Total Corpus Quality. *Socius*, *8*, 1–15. https://doi.org/10.1177/23780231221135523

Hwang, J., & Naik, N. (2023). Systematic Social Observation at Scale: Using Crowdsourcing and Computer Vision to Measure Visible Neighborhood Conditions. *Sociological Methodology*, *53*(2), 183–216. https://doi.org/10.1177/00811750231160781

Jeong, W., & Han, H. (2015). Media Visualization of Book Cover Images: Exploring Differences Among Bestsellers in Different Countries. *Proceedings of the Association for Information Science and Technology*, *52*(1), 1–4. https://doi.org/10.1002/pra2.2015.1450520100107

Kitchin, R. (2016). Thinking Critically About and Researching Algorithms. *Information, Communication & Society*, *20*(1), 14–29. https://doi.org/10.1080/1369118X.2016.1154087

Kitchin, R. (2025). *Critical Data Studies: An A to Z Guide to Concepts and Methods*. Polity Press.

Lowe, D. G. (2004). Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, *60*(2), 91–110. https://doi.org/10.1023/B:VISI.0000029664.99615.94

Manovich, L. (2012a). How to Compare One Million Images? In D. M. Berry (Hrsg.), *Understanding Digital Humanities* (S. 249–278). Palgrave Macmillan. https://doi.org/10.1057/9780230371934_14

Manovich, L. (2012b). Media visualization. In A. N. Valdivia (Hrsg.), *The International Encyclopedia of Media Studies* (S. 1–21). John Wiley & Sons. https://doi.org/10.1002/9781444361506.wbiems144

Manovich, L. (2020). *Cultural Analytics*. MIT Press. https://doi.org/10.7551/mitpress/11214.001.0001

Marguin, S., Pelger, D., & Stollmann, J. (2024). Mappings as Joint Spatial Display. In A. J. Heinrich, S. Marguin, A. Million, & J. Stollmann (Hrsg.), *Handbook of Qualitative and Visual Methods in Spatial Research* (S. 295–312). transcript. https://doi.org/10.1515/9783839467343-023

Mcilvenny, P. B., & Davidsen, J. (2017). A Big Video Manifesto: Re-sensing Video and Audio. *Nordicom Information*, *39*(2), 15–21.

Mitchell, W. J. T. (1994). Metapictures. In *Picture Theory: Essays on Verbal and Visual Representation* (S. 35–82). University of Chicago Press.

MoMA. o. J. Collection. *GitHub*. https://github.com/MuseumofModernArt/collection (letzter Zugriff am 30. Oktober 2024).

Murphy, O., Villaespesa, E., Bernhardt, J., Golgath, T., & Thiel, S. (2022). *Künstliche Intelligenz und Museen: Ein Toolkit*. Goldsmiths / University of London. https://doi.org/10.17613/abet-w606

Rogers, R. (2021). Visual Media Analysis for Instagram and Other Online Platforms. *Big Data & Society*, *8*(1), 1–23. https://doi.org/10.1177/20539517211022370

Rösch, A. (2022). *Algorithmische Klassifikation: Ein nützliches Instrument für digitale Kopfschmerztagebücher in mHealth Apps?* Dissertation, Charité / Universitätsmedizin Berlin. http://dx.doi.org/10.17169/refubium-34728

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (S. 4510–4520). IEEE. https://doi.org/10.1109/CVPR.2018.00474

Schranz, C. (Hrsg.). (2021). *Shifts in Mapping: Maps as a Tool of Knowledge*. transcript. https://doi.org/10.1515/9783839460412

Shen, D., Shen, H., Zhu, H., & Marron, J. S. (2012). High Dimensional Principal Component Scores and Data Visualization. *arXiv*. Preprint / Working Paper. https://doi.org/10.48550/arXiv.1211.2679

TED Conferences. o. J. The Best Hans Rosling Talks You've Ever Seen [Playlist]. *TED*. https://www.ted.com/playlists/474/the_best_hans_rosling_talks_yo (letzter Zugriff am 30. Oktober 2024).

Tschuggnall, M., Specht, G., & Riepl, C. (2016). Algorithmisch unterstützte Literarkritik: Eine grammatikalische Analyse zur Bestimmung von Schreibstilen. In H. Rechenmacher (Hrsg.), *Arbeiten zu Text und Sprache im Alten Testament* (S. 415–428). EOS.

Unternährer, M. (2024). *Momente der Datafizierung: Zur Produktionsweise von Personendaten in der Datenökonomie*. transcript. https://doi.org/10.1515/9783839470596

van der Maaten, L., & Hinton, G. (2008). Visualizing Data Using t-SNE. *Journal of Machine Learning Research*, *9*(86), 2579–2605.

Wattenberg, M., Viégas, F., & Johnson, I. (2016). How to Use t-SNE Effectively. *Distill*. https://distill.pub/2016/misread-tsne/ (letzter Zugriff am 30. Oktober 2024).

Windhager, F., Federico, P., Schreder, G., Glinka, K., Dörk, M., Miksch, S., & Mayr, E. (2019). Visualization of Cultural Heritage Collection Data: State of the Art and Future Challenges. *IEEE Transactions on Visualization and Computer Graphics*, *25*(6), 2311–2330. https://doi.org/10.1109/TVCG.2018.2830759