

Thomas Bartz-Beielstein
Eva Bartz *Hrsg.*

Online Machine Learning

Eine praxisorientierte Einführung

 Springer Vieweg

Online Machine Learning

Thomas Bartz-Beielstein · Eva Bartz
(Hrsg.)

Online Machine Learning

Eine praxisorientierte Einführung

 Springer Vieweg

Hrsg.

Thomas Bartz-Beielstein
Institute for Data Science, Engineering,
and Analytics
TH Köln
Gummersbach, Deutschland

Eva Bartz
Bartz & Bartz GmbH
Gummersbach, Deutschland

ISBN 978-3-658-42504-3 ISBN 978-3-658-42505-0 (eBook)
<https://doi.org/10.1007/978-3-658-42505-0>

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <https://portal.dnb.de> abrufbar.

© Der/die Herausgeber bzw. der/die Autor(en), exklusiv lizenziert an Springer Fachmedien Wiesbaden GmbH, ein Teil von Springer Nature 2024

Das Werk einschließlich aller seiner Teile ist urheberrechtlich geschützt. Jede Verwertung, die nicht ausdrücklich vom Urheberrechtsgesetz zugelassen ist, bedarf der vorherigen Zustimmung des Verlags. Das gilt insbesondere für Vervielfältigungen, Bearbeitungen, Mikroverfilmungen und die Einspeicherung und Verarbeitung in elektronischen Systemen.

Die Wiedergabe von allgemein beschreibenden Bezeichnungen, Marken, Unternehmensnamen etc. in diesem Werk bedeutet nicht, dass diese frei durch jedermann benutzt werden dürfen. Die Berechtigung zur Benutzung unterliegt, auch ohne gesonderten Hinweis hierzu, den Regeln des Markenrechts. Die Rechte des jeweiligen Zeicheninhabers sind zu beachten.

Der Verlag, die Autoren und die Herausgeber gehen davon aus, dass die Angaben und Informationen in diesem Werk zum Zeitpunkt der Veröffentlichung vollständig und korrekt sind. Weder der Verlag noch die Autoren oder die Herausgeber übernehmen, ausdrücklich oder implizit, Gewähr für den Inhalt des Werkes, etwaige Fehler oder Äußerungen. Der Verlag bleibt im Hinblick auf geografische Zuordnungen und Gebietsbezeichnungen in veröffentlichten Karten und Institutionsadressen neutral.

Planung/Lektorat: David Imgrund

Springer Vieweg ist ein Imprint der eingetragenen Gesellschaft Springer Fachmedien Wiesbaden GmbH und ist ein Teil von Springer Nature.

Die Anschrift der Gesellschaft ist: Abraham-Lincoln-Str. 46, 65189 Wiesbaden, Germany

Das Papier dieses Produkts ist recycelbar.

Vorwort

Dieses Buch beschäftigt sich mit dem spannenden, zukunftssträchtigen Thema des Online Machine Learning (OML). Es gliedert sich in drei Teile:

Zunächst beschäftigen wir uns ausführlich mit den theoretischen Grundlagen von OML. Wir beschreiben, was OML ist, und fragen, wie man es mit Batch Machine Learning (BML) vergleichen kann und welche Kriterien für einen aussagekräftigen Vergleich man entwickeln sollte. Im zweiten Teil stellen wir Überlegungen zur Praxis an und belegen diese im dritten Teil mit konkreten praktischen Anwendungen.

Warum OML? Es geht unter anderem um den entscheidenden Zeitvorteil. Das können Monate, Wochen, Tage, Stunden oder auch nur Sekunden sein. Dieser Zeitvorteil kann entstehen, wenn die Künstliche Intelligenz (KI) Daten fortlaufend, also online auswerten kann. Sie nicht darauf warten muss, bis ein kompletter Datensatz zur Verfügung steht, sondern bereits eine einzelne Beobachtung zur Aktualisierung des Modells verwendet werden kann. Hat OML noch andere Vorteile außer dem offensichtlichen Zeitvorteil? Wenn ja, welche? Wir fragen: Gibt es Grenzen des BML, die OML überwindet?

Es muss genau untersucht werden, zu welchem Preis man sich diese Vorteile durch OML verschafft. Wie hoch ist der Speicherbedarf im Vergleich zu herkömmlichen Verfahren? Speicherbedarf bedeutet auch finanzielle Kosten, z. B. durch höheren Energiebedarf. Ist OML eventuell energiesparend und damit nachhaltiger, also Green IT? Erhält man vergleichbar gute Ergebnisse? Leidet die Güte (Performanz), werden die Ergebnisse ungenauer? Um diese Fragen verlässlich zu beantworten, geben wir im Theorieteil zunächst eine verständliche Einführung in OML, die sich sowohl für Anfänger als auch für Fortgeschrittene eignet. Dann begründen wir die von uns gefundenen Kriterien, die wir für die Vergleichbarkeit von OML und BML heranziehen, nämlich eine gut nachvollziehbare Darstellung von Güte, Zeit- und Speicherbedarf.

Im zweiten Teil beschäftigen wir uns mit der Frage, genau wie OML in der Praxis eingesetzt werden kann. Zu Wort kommen Experten aus der Praxis, die von den Anforderungen an die amtliche Statistik berichten. Wir begründen Empfehlungen für den praktischen Einsatz von OML. Wir stellen umfassend die Softwarepakete vor, die derzeit

für OML zur Verfügung stehen, insbesondere „river“¹, und bieten mit Sequential Parameter Optimization Toolbox for River (spotRiver) eine von uns eigens für OML entwickelte Software an. Wir beschäftigen uns ausführlich mit den besonderen Problemen, die bei Datenströmen auftreten können. Hier sei das für Datenströme zentrale Problem der Drift genannt. Wir behandeln die Erklärbarkeit von KI-Modellen, die Interpretierbarkeit und Reproduzierbarkeit. Diese Aspekte können zu höherer Akzeptanz von KI beitragen, wie sie in kommenden Regulierungen für KI-Systeme gefordert wird.

Im Anwendungsteil präsentieren wir zwei ausführliche Studien, davon eine mit einem großen Datensatz mit einer Million Daten. Wir belegen, wann OML besser funktioniert als BML. Besonders interessant ist die Studie zum Hyperparameter-Tuning von OML. Hier zeigen wir, wie OML durch die Optimierung von Hyperparametern deutlich besser funktionieren kann.

Notebooks

Ergänzender Programmcode zu den Anwendungen und Beispielen aus diesem Buch ist in sogenannten „Jupyter-Notebooks“ im GitHub-Repository <https://github.com/sn-code-inside/online-machine-learning> zu finden. Die Notebooks sind kapitelweise organisiert.

Die Unternehmensberatung Bartz & Bartz GmbH² hat den Grundstein für dieses Buch gelegt, als sie 2022 einen Auftrag aus einer Ausschreibung des Statistischen Bundesamtes zugeschlagen bekam³. Das Statistische Bundesamt wollte wissen, ob es für den Schatz an Daten und die Auswertung im öffentlichen Auftrag Sinn macht, OML jetzt schon einzusetzen (siehe hierzu die Ausführungen in Kap. 7). Unser leicht ernüchterndes Ergebnis der Expertise war: Es eröffnen sich interessante Perspektiven für die Zukunft, aber momentan bietet sich ein Einsatz noch nicht unmittelbar an. Teils gibt es in der Praxis fachliche und organisatorische Hürden, Prozesse so anzupassen, dass die Vorteile von OML wirklich zur Geltung kommen können. Teils sind OML-Verfahren und Implementierungen noch nicht ausgereift genug.

Das Thema hat uns so fasziniert, dass wir uns entschlossen haben, es weiter zu verfolgen. Prof. Dr. Thomas Bartz-Beielstein hat die Frage nach der Praxisrelevanz von OML mit in die TH Köln genommen und dort seine seit Jahren laufende Forschung zu dem Bereich weiter voran getrieben. Die Forschergruppe am Institut für Data Science, Engineering, and Analytics (IDE+A)⁴ konnte unter seiner Anleitung Software so weit entwickeln, dass wir glauben, die Tauglichkeit ein ganzes Stück voran gebracht zu haben. So haben

¹ <https://riverml.xyz/>

² <https://bartzundbartz.de>

³ <https://destatis.de>

⁴ <https://www.th-koeln.de/idea>

wir die damals entstandene Expertise der Bartz & Bartz GmbH mit der Forschung an der TH Köln kombiniert, woraus dieses Buch entstanden ist.

Insgesamt eignet sich das Buch gleichermaßen als Handbuch zum Nachschlagen für Experten, die sich mit OML beschäftigen, als Lehrbuch für Anfänger, die sich mit OML beschäftigen wollen, und als wissenschaftliche Publikation für Wissenschaftler, die sich mit OML beschäftigen, da es den neuesten Stand der Forschung wiedergibt. Es kann aber auch quasi als OML-Consulting dienen, denn Entscheider und Praktiker können anhand unserer Ausführungen OML maßgeschneidert für ihre Bedürfnisse anpassen, für ihre Anwendung einsetzen und fragen, ob die Vorteile von OML eventuell die Kosten aufwiegen.

Um nur einige Beispiele aus der militärischen und zivilen Praxis zu nennen:

- Sie verwenden hochmoderne Sensorsysteme, um Hochwasser vorauszusagen. Hier kann eine schnellere Vorhersage Leben retten.
- Sie müssen terroristische Angriffe abwehren und setzen dazu Unterwassersensorik ein. Hier kann es entscheidend sein, dass die KI schneller „erkennt“, ob es sich um harmlose Wassersportler handelt.
- Sie sind verantwortlich für die Beobachtung des Luftraums. Aufklärungsdrohnen können beispielsweise effizienter eingesetzt werden, wenn sie mit ganz aktuellen KI-Datenauswertungen programmiert und trainiert werden können.
- Sie müssen sehr zügig die Produktion von Gütern der kritischen Infrastruktur, wie Impfstoff, Schutzkleidung oder medizinische Apparaturen, anpassen. Hier kann es sinnvoll sein, den gesamten Produktionsablauf samt einzusetzender Rohstoffe so aktuell wie möglich zu gestalten. Dazu kann eine Echtzeitauswertung und Übersetzung in Bedarfe anhand der Krankenhausbettenbelegung oder Krankenschreibungen dienen.
- Sie müssen als Zahlungsdienstleister Betrugsversuche quasi in Echtzeit erkennen.

Abschließend stellen wir fest: OML wird bald praxistauglich, es lohnt sich, sich jetzt schon damit zu beschäftigen. In diesem Buch werden schon einige Werkzeuge vorgestellt, die in Zukunft die Praxis von OML erleichtern werden. Ein vielversprechender Durchbruch steht zu erwarten, weil die Praxis zeigt, dass aufgrund der großen Datenmengen, die in der Praxis anfallen, das bisherige BML nicht mehr ausreicht. OML ist die Lösung, um die Datenströme in Echtzeit auszuwerten, zu verarbeiten und Ergebnisse zu liefern, die für die Praxis relevant sind.

Im Einzelnen behandelt das Buch folgende Themen: Kapitel 1 beschreibt die Motivation für dieses Buch und die Zielsetzung. Es beschreibt die Nachteile und Grenzen von BML und die Notwendigkeit für OML. Kapitel 2 gibt eine Übersicht und Bewertung von Verfahren und Algorithmen mit speziellem Fokus auf Supervised Learning (Klassifikation und Regression). Kapitel 3 beschreibt Verfahren zur Drifterkennung. Die Aktualisierbarkeit der OML-Verfahren wird in Kap. 4 behandelt. Kapitel 5 erläutert Verfahren zur Bewertung von OML-Verfahren. Kapitel 6 beschäftigt sich mit den besonderen

Anforderungen an OML. Mögliche OML-Anwendungen werden in Kap. 7 dargestellt und von Experten der amtlichen Statistik beurteilt. Die Verfügbarkeit der Algorithmen in Softwarepaketen, im Speziellen für R und Python, wird in Kap. 8 dargestellt.

Der benötigte Rechenaufwand bei der Aktualisierung der OML-Modelle, auch im Vergleich zu einem algorithmisch ähnlichen Offline-Verfahren (BML), wird experimentell in Kap. 9 untersucht. Dort wird auch darauf eingegangen, inwiefern die Modellgüte beeinträchtigt werden könnte, insbesondere im Vergleich zu ähnlichen Offline-Verfahren. Kapitel 10 beschreibt das Hyperparameter-Tuning für OML. Kapitel 11 präsentiert eine Zusammenfassung und gibt wichtige Empfehlungen für die Praxis.

Gummersbach
April 2023

Eva Bartz

Inhaltsverzeichnis

1	Einleitung: Vom Batch Machine Learning zum Online Machine Learning	1
	Thomas Bartz-Beielstein	
1.1	Datenströme	2
1.2	Nachteile des Batch-Lernens	3
1.2.1	Speicherbedarf	4
1.2.2	Drift	4
1.2.3	Neue, unbekannte Daten	8
1.2.4	Zugänglichkeit und Verfügbarkeit der Daten	9
1.2.5	Weitere Probleme	9
1.3	Inkrementelles Lernen, Online-Lernen und Stream-Lernen	10
1.4	Überführung des Batch Machine Learning in das Online Machine Learning	12
2	Supervised Learning: Klassifikation und Regression	13
	Thomas Bartz-Beielstein	
2.1	Klassifikation	14
2.1.1	Baselinealgorithmen	14
2.1.2	Der naive Bayes-Klassifikator	14
2.1.3	Baumbasierte Verfahren	16
2.1.4	Weitere Klassifikationsverfahren	19
2.2	Regression	20
2.2.1	Online Linear Regression	20
2.2.2	Hoeffding Tree Regressor	20
2.3	Ensemble-Methoden für Online Machine Learning	21
2.4	Clustering	21
2.5	Übersicht: Online Machine Learning-Verfahren	22
3	Drifterkennung und -behandlung	23
	Thomas Bartz-Beielstein	
3.1	Architekturen für Driftbehandlungsmethoden	24

3.1.1	Adaptive Schätzer	24
3.1.2	Change Detectors	25
3.1.3	Ensemblebasierte Ansätze	25
3.2	Grundlegende Überlegungen zu Fenstertechniken	26
3.3	Populäre Verfahren zur Drifterkennung	27
3.3.1	Statistische Tests zur Drifterkennung und Change Detection	27
3.3.2	Kontrollkarten (Drift Detection Method)	27
3.3.3	Adaptive Windowing	28
3.3.4	Implizite Drifterkennungsalgorithmen	30
3.4	Online Machine Learning Algorithmen mit Drifterkennung: Hoeffding Window Trees	31
3.4.1	Concept-adapting Very Fast Decision Trees	31
3.4.2	Hoeffding Adaptive Trees	33
3.4.3	Übersicht: Hoeffding Window Trees	33
3.4.4	Übersicht: HT in river	33
4	Initiale Auswahl und nachträgliche Aktualisierung von OML-Modellen	37
	Thomas Bartz-Beielstein	
4.1	Initiale Modellauswahl	38
4.2	Modelländerungen	39
4.2.1	Hinzufügen neuer Merkmale	39
4.2.2	Manuelle Modelländerungen als Reaktion auf Drift	39
4.2.3	Sicherstellung der Modellgüte nach einem Modellupdate	40
4.3	Katastrophales Vergessen	40
4.3.1	Definition: Katastrophales Vergessen	41
4.3.2	Methoden gegen das katastrophale Vergessen	41
5	Evaluation und Performanzmessung	43
	Thomas Bartz-Beielstein	
5.1	Auswahlmethode	44
5.1.1	Holdout	44
5.1.2	Progressive Validierung: Interleaved test-then-train	45
5.1.3	Maschinelles Lernen im Batch-Verfahren mit einem Vorhersagehorizont	46
5.1.4	Landmark Batch Machine Learning mit einem Vorhersagehorizont	47
5.1.5	Window Batch Machine Learning mit einem Vorhersagehorizont	47
5.1.6	Online Machine Learning mit einem Vorhersagehorizont	48
5.1.7	Online-Maschinelles Lernen	48

5.2	Bestimmung des Training- und Testdatensatzes im Paket <code>spotRiver</code>	50
5.2.1	Methoden für Batch Machine Learning und Online Machine Learning	50
5.2.2	Methoden für Online Machine Learning (River)	52
5.3	Performanz des Algorithmus/Modells	56
5.4	Datenstrom- und Driftgeneratoren	58
5.4.1	Data-Stream-Generatoren in den scikit-Paketen	58
5.4.2	SEA-Drift-Generator	58
5.4.3	Friedman-Drift-Generator	59
5.5	Zusammenfassung	59
6	Besondere Anforderungen an OML-Verfahren	61
	Thomas Bartz-Beielstein	
6.1	Fehlende Daten, Imputation	62
6.2	Kategorische Attribute	63
6.3	Ausreißer	63
6.3.1	Weitere Anomalieerkennungsverfahren für Zeitreihen	64
6.3.2	One-Class Support Vector Machine zur Anomalieerkennung	64
6.3.3	Verfügbare Algorithmen zur Anomalieerkennung in <code>river</code>	64
6.4	Unbalancierte (unausgewogene) Daten	65
6.5	Große Anzahl an Features (Attributen)	65
6.6	FAIR, Interpretierbarkeit und Erklärbarkeit	66
7	Praxisanwendungen	69
	Steffen Moritz, Florian Dumpert, Thomas Bartz-Beielstein und Eva Bartz	
7.1	Anwendungen und Anwendungsperspektiven in der amtlichen Statistik	70
7.1.1	Potenziale und Herausforderungen	71
7.1.2	Vereinbarkeit mit Qualitätskriterien	76
7.1.3	Einbettung in den Statistikproduktionsprozess	78
7.1.4	(Online-)Machine-Learning-Anwendungen in Statistikinstitutionen	79
7.2	Andere Anwendungen mit Bezug zur amtlichen Statistik	81
7.2.1	Immobilienpreise	82
7.2.2	Pandemie-Vorhersagen	82
7.2.3	Stimmungsvorhersagen für Wahlen	83
7.2.4	Nowcasting für Wirtschaftsindizes	84
7.3	Aspekte bezüglich des Praxiseinsatzes	85

7.3.1	Unterschiede im Deployment-Prozess bei Batch Machine Learning- und Online Machine Learning-Ansätzen	85
7.3.2	Personalaufwand	86
8	Open-Source-Software für Online Machine Learning	89
	Thomas Bartz-Beielstein	
8.1	Übersicht und Beschreibung der Softwarepakete	90
8.1.1	Massive Online Analysis	90
8.1.2	Massive Online Analysis in R	90
8.1.3	stream	91
8.1.4	river	91
8.2	Softwareumfang	93
8.3	Vergleich der Programmiersprachen	95
9	Ein experimenteller Vergleich von Batch- und Online-Machine-Learning-Algorithmen	97
	Thomas Bartz-Beielstein	
9.1	Studie: Bike Sharing	98
9.1.1	Modellübersicht	101
9.1.2	Lineare Regression	101
9.1.3	Gradient Boosting	105
9.1.4	Hoeffding-Regressionsbäume	108
9.1.5	Abschließender Vergleich der Bike-Sharing-Experimente	110
9.1.6	Zusammenfassung: Bike-Sharing-Experimente	110
9.2	Studie: Sehr große Datensätze mit Drift	112
9.2.1	Der Friedman-Drift-Datensatz	112
9.2.2	Algorithmen	113
9.2.3	Ergebnisse	113
9.3	Zusammenfassung	114
10	Hyperparameter-Tuning	117
	Thomas Bartz-Beielstein	
10.1	Hyperparameter-Tuning: Eine Einführung	118
10.2	Die Hyperparameter-Tuning-Software Sequential Parameter Optimization Toolbox	119
10.3	Studie: Hyperparameter-Tuning des Hoeffding Adaptive Tree Regressor-Algorithmus auf den Friedman-Drift-Daten	120
10.3.1	Laden der Daten	120
10.3.2	Spezifikation des Vorverarbeitungsmodells	121
10.3.3	Auswahl des zu tunenden Algorithmus und der Defaulthyperparameter	122
10.3.4	Modifikation der Defaultwerte für die Hyperparameter	122
10.3.5	Auswahl der Zielfunktion (Loss-Funktion)	124

10.3.6 Aufruf des Hyperparameter Tuners Sequential Parameter Optimization Toolbox	125
10.3.7 Ergebnisse des Hoeffding Adaptive Tree Regressor-Tunings	125
10.3.8 Erklärbarkeit und Verständnis	128
10.4 Zusammenfassung	132
11 Zusammenfassung und Ausblick	133
Thomas Bartz-Beielstein und Eva Bartz	
11.1 Notwendigkeit für OML-Verfahren	133
11.2 Empfehlungen für die Online Machine Learning-Praxis	134
A Definitionen und Erläuterungen	137
A.1 Gradientenabstieg	137
A.2 Satz von Bayes	137
A.3 Hoeffding-Schranke	138
A.4 Kappa-Statistiken	138
Zusatzmaterial	141
B.1 Notebooks	141
B.2 Software	142
Glossary	143
Literatur	145
Stichwortverzeichnis	151

Autorenverzeichnis

Thomas Bartz-Beielstein Institute for Data Science, Engineering, and Analytics, TH Köln, Gummersbach, Deutschland

Eva Bartz Bartz & Bartz GmbH, Gummersbach, Deutschland

Florian Dumpert Statistisches Bundesamt, Wiesbaden, Deutschland

Steffen Moritz Statistisches Bundesamt, Wiesbaden, Deutschland



Einleitung: Vom Batch Machine Learning zum Online Machine Learning

1

Thomas Bartz-Beielstein

Inhaltsverzeichnis

1.1	Datenströme	2
1.2	Nachteile des Batch-Lernens	3
1.2.1	Speicherbedarf	4
1.2.2	Drift	4
1.2.3	Neue, unbekannte Daten	8
1.2.4	Zugänglichkeit und Verfügbarkeit der Daten	9
1.2.5	Weitere Probleme.....	9
1.3	Inkrementelles Lernen, Online-Lernen und Stream-Lernen	10
1.4	Überführung des Batch Machine Learning in das Online Machine Learning	12

Zusammenfassung

Batch Machine Learning (BML), das auch als „Offline Machine Learning“ bezeichnet wird, stößt bei sehr großen Datenmengen an seine Grenzen. Dies betrifft insbesondere den verfügbaren Speicher, das Behandeln von Drift in Datenströmen und die Verarbeitung neuer, unbekannter Daten. Online Machine Learning (OML) ist eine Alternative zu BML, die die Grenzen von BML überwindet. In diesem Kapitel werden die grundlegenden Begriffe und Konzepte von OML vorgestellt, wodurch die Unterschiede zum BML sichtbar werden.

T. Bartz-Beielstein (✉)

Institute for Data Science, Engineering, and Analytics, TH Köln, Gummersbach, Deutschland

E-Mail: thomas.bartz-beielstein@th-koeln.de

© Der/die Autor(en), exklusiv lizenziert an Springer Fachmedien Wiesbaden GmbH, ein Teil von Springer Nature 2024

T. Bartz-Beielstein und E. Bartz (Hrsg.), *Online Machine Learning*,

https://doi.org/10.1007/978-3-658-42505-0_1

1

1.1 Datenströme

Das Volumen der aus verschiedenen Quellen generierten Daten hat in den letzten Jahren enorm zugenommen. Technologische Fortschritte haben die kontinuierliche Erfassung von Daten ermöglicht. Zur Beschreibung von Big Data wurden anfänglich die „drei Vs“ (Volume, Velocity und Variety) als Kriterien verwendet¹: Volume bezeichnet hierbei die große Menge an Daten, Velocity die hohe Geschwindigkeit, mit der die Daten generiert werden, und Variety die große Vielfalt der Daten.

Die in diesem Buch betrachteten Datenströme (Streamingdaten) stellen eine noch größere Herausforderung für Machine Learning (ML)-Algorithmen dar als Big Data. Zu den drei Big-Data-Vs kommen noch weitere Herausforderungen hinzu, die sich insbesondere aus der Flüchtigkeit (Volatilität) und der Möglichkeit, dass abrupte Änderungen („Drift“) auftreten können, ergeben.

Definition 1.1 (Streamingdaten)

Streamingdaten sind Daten, die in einem kontinuierlichen Datenstrom erzeugt werden. Sie sind lose strukturiert, flüchtig (nur einmal verfügbar), immer „fließend“ und beinhalten unvorhersehbare, teilweise abrupte, Änderungen. Streamingdaten sind eine Teilmenge von Big Data mit den folgenden Eigenschaften:

- Volumen: Streamingdaten werden in sehr großen Mengen erzeugt.
- Velocity: Streamingdaten werden in sehr hoher Geschwindigkeit erzeugt.
- Variety: Streamingdaten sind in sehr unterschiedlichen Formaten verfügbar. Diese Eigenschaft bezeichnen wir als „vertikale Vielfalt“.
- Variability: Streamingdaten sind strukturlos und variieren im Laufe der Zeit. Beispielsweise kann graduell oder abrupt Drift auftreten. Diese Eigenschaft bezeichnen wir als „horizontale Vielfalt“.
- Volatilität: Streamingdaten sind flüchtig und nur einmal verfügbar.

Beispiel Streamingdaten

Bei verschiedenen täglichen Transaktionen, z. B. beim Onlineshopping, beim Onlinebanking oder beim Onlinehandel mit Aktien, werden sehr viele Daten erzeugt. Hinzu kommen Sensordaten, Social-Media-Daten, Daten aus Betriebsüberwachungen und Daten aus dem Internet der Dinge, um nur einige Beispiele zu nennen. ◀

Streamingdaten erfordern Analysen in Echtzeit oder nahezu in Echtzeit. Da der Datenstrom ständig produziert wird und nie endet, ist es nicht möglich, diese enormen Datenmengen

¹ Die drei Vs wurden im Laufe der Zeit durch Hinzunahme von Veracity und Value zu den „fünf Vs“ erweitert.