

4 Datenvisualisierung

PROF. DR. SARA IRINA FABRIKANT

Studium Digitale Kursbaustein Effektive Informationsvisualisierung

Begleit-Skript

Dieses Skript bereitet die Inhalte des Kursbaustein Effektive Informationsvisualisierung schriftlich auf. Die Inhalte sind nicht identisch mit den Videos; sie wurden angepasst für die Lesbarkeit als Dokument.

4.1 Lektion 1: Einstieg und Überblick

Die Lernziele dieses Kursbausteins sind:

- Sie erkennen, wann es sich lohnt, Ihre Zeit in eine Visualisierung zu investieren.
- Sie kennen diverse Visualisierungsmethoden und können diese zielgerecht einsetzen.
- Sie können die für Ihre Aufgaben und Daten angemessene Visualisierungsmethode auswählen und anwenden.
- Sie können selbständig attraktive Visualisierungen mit Microsoft Excel erstellen.

Als Orientierungshilfe ist der Kursbaustein in die 5 Ws der Visualisierung gegliedert: Warum, Was, für Wen, Wozu und Wie wird visualisiert?

Eine Visualisierung macht dann Sinn, wenn damit auf einen Blick aufgenommen werden kann, was sonst in vielen Worten beschrieben werden müsste. Bei der Wahl der Visualisierungsmethode ist es wichtig, für welche Zielgruppe sie bestimmt ist und wie diese sie wahrnimmt.

Zuerst gilt es abzuklären, welche Fragen die Visualisierung beantworten soll und welche Daten dafür zur Verfügung stehen. Erst danach kann eine passende Visualisierungsform gewählt werden. Häufig werden Mengen, Verteilungen, Anteile und Zusammenhänge visualisiert. Zeitreihen und Daten mit Raumbezug sind ebenfalls Gegenstand von Visualisierungen, auf diese wird hier aber nicht genauer eingegangen.

Zum begleiteten Selbststudium werden die folgenden Bücher empfohlen:

- Claus Wilke, <https://serialmentor.com/dataviz/>
- Stephanie Evergreen, <https://stephanieevergreen.com/books/>

4.2 Lektion 2: Warum will ich visualisieren?

Diese Lektion hat folgende Lernziele:

- Sie können informiert entscheiden, wann es sich lohnt, Daten zu visualisieren.
- Sie können selbständig eine statistische Aussage mit einem Bild visualisieren.
- Sie können mit einem guten Beispiel illustrieren, warum die Visualisierung von Daten wichtig und zielführend ist.

Ein wichtiger Grund für die Visualisierung ist wenn die Kommunikation eines Sachverhalts durch ein Bild prägnanter möglich ist, als durch Text. Interessante Muster in den Daten können so viel effizienter erkannt werden, als wenn die Daten in Rohform vorliegen. Dies ist auch der Grund dafür, dass Daten vor statistischen Analysen zuerst grafisch dargestellt werden sollten. Sind die Daten bekannt, beantwortet das erhaltene Datenmuster den Sachverhalt der kommuniziert werden soll. Müssen die Daten zuerst weiter erkundet werden, wirft das visualisierte Muster weiterführende Fragen oder zu testende Hypothesen auf.

Die Visualisierung kann genutzt werden, um real existierende Muster von Scheinmustern in Datenhaufen zu unterscheiden. Dafür muss das Muster durch die grafischen Mittel auch unmissverständlich wahrgenommen werden können.

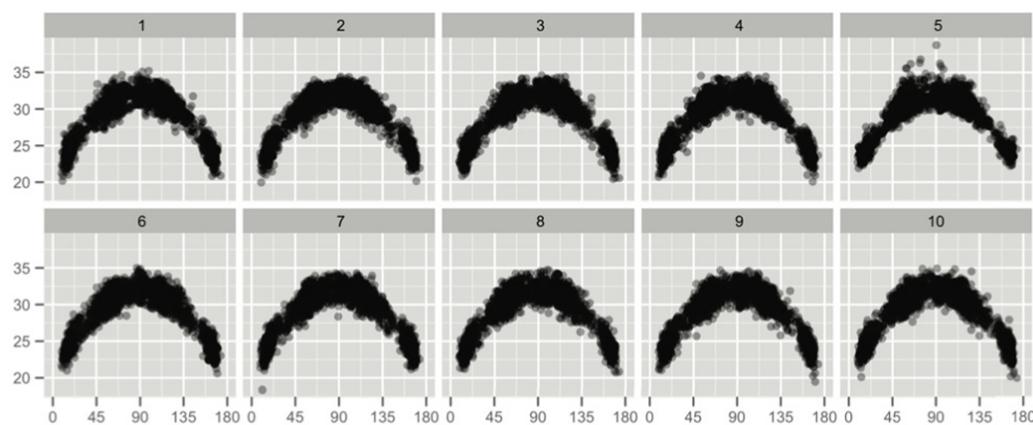


Abbildung aus: Wickham, H., Cook, D., Hofmann, H.. Buja, A. (2010). Graphical Inference for Infovis. IEEE Transactions on Visualization and Computer Graphics, vol. 16, no. 6: 973-979.

4.3 Lektion 3: Was will ich visualisieren?

Diese Lektion hat folgende Lernziele:

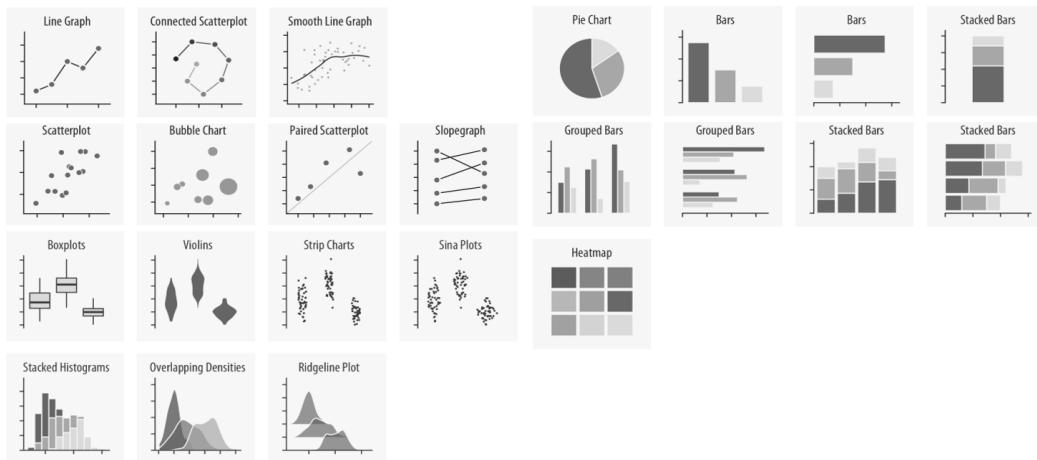
- Sie kennen zentrale Dateneigenschaften (Datentyp, Datenmerkmale, Messniveau, etc.).
- Sie kennen die graphischen Variablen für die Visualisierung.
- Sie haben einen Überblick über das Visualisierungsinventar.

Es gibt qualitative Daten, wie z.B. Texte, die wir unstrukturierte Daten nennen, da nicht in eine Datentabelle passen. Quantitative Daten befinden sich bereits aufbereitet in einer Datentabelle. Bei diesen Daten geht es in einem nächsten Schritt darum, ihre Datentypen und -merkmale zu identifizieren. Diskrete Daten z.B. die Anzahl Einwohner eines Landes und kontinuierliche Daten der durchschnittliche Flächenverbrauch eines Golfplatzes. Dieses Wissen ist wichtig, um richtig zu visualisieren.

Typ	Merkmal	Wert	Messniveau	Beschreibung
Numerisch	kontinuierlich	1.3, 10 ²	ratio	Distanz in Metern
Numerisch	diskret	1, 10, -25	interval	Temperatur in Grad Celsius
alphabetisch	diskret	niedrig, mittel, hoch	ordinal	Erfahrung in Visualisierung
alphabetisch	diskret	Rot, Grün, Blau	nominal	Farbton
Alpha-numerisch	Datum/Zeit	1. August 1291	interval/ratio	Nationalfeiertag

Tabelle 1: Metadaten einer Datentabelle

Für die Darstellung von Daten stehen diverse geometrische Möglichkeiten zur Verfügung. Diese Formen sind nach geometrischen Grundformen (Punkte, Linien und Flächen) sortiert.



Neben der Geometrie steht auch ein Inventar an graphischen Variablen zur Verfügung. Dazu zählen:

- die Lage der Datenpunkte
- die Orientierung der Linien
- der Farbton
- die Helligkeit einer Farbe
- deren Sättigung
- die Textur
- die Form einer Linie oder Fläche

4.4 Lektion 4: Für wen wird visualisiert?

Die Lernziele dieser Lektion sind:

- Sie kennen typische visuelle Wahrnehmungseffekte.
- Sie wissen um das Prinzip der graphischen Prägnanz
- Sie können die Bausteine des Gesetzes der guten Gestalt nennen.

- Sie können die Gestaltbausteine selbstständig in Visualisierungen anwenden.

Genau wie Texte, Musik oder Filme haben Visualisierungen ein Zielpublikum. Der Kontext der Nutzung und das Vorwissen der Personen sind dabei von grosser Bedeutung.

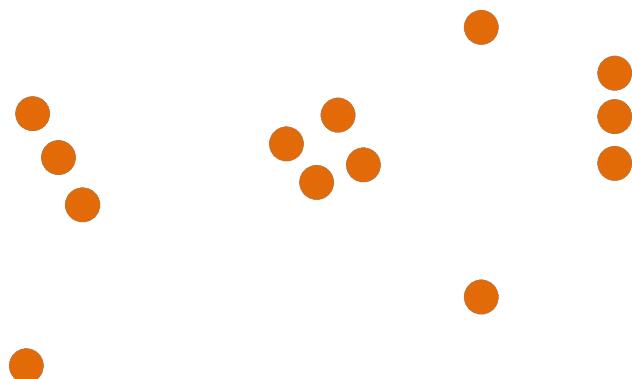
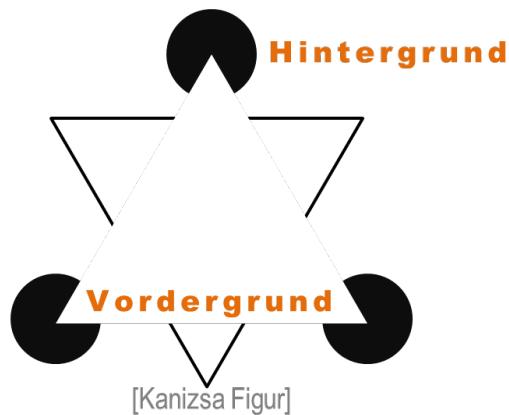
Die menschliche Wahrnehmung ist nicht perfekt und lässt sich täuschen. Wir sind keine physikalischen Lichtsensoren, sondern können mithilfe unserer Erfahrung und des Nutzungskontexts visuelle Eindrücke besser einordnen und daraus sinnvolle Informationen gewinnen.

Das Vorlesen der Farbnamen, die in jeweils anderen Farben geschrieben sind, fällt schwer, da der Inhalt der Wörter nicht zum visuellen Input passt. Wir sprechen dabei von einem Konflikt zwischen einem bottom-up- und einem top-down-Prozess. Der bottom-up-Prozess liefert die sensorische, visuelle Information des Sehsinns, während der top-down-Prozess der Wissensverarbeitung gleichzeitig widersprüchliche Informationen zum Inhalt der Wörter liefert. Was Leute sehen, hängt von ihrem Vorwissen und ihrer Erfahrung ab.

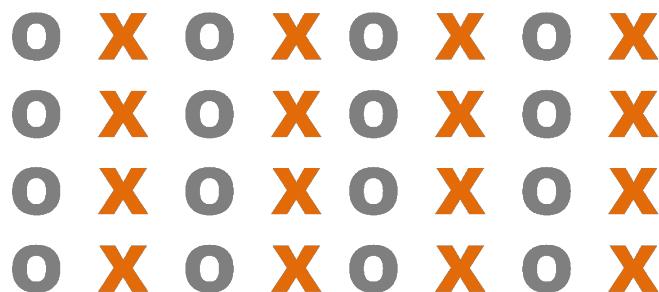
In der Visualisierung sollen zweideutige Darstellungen um jeden Preis verhindert werden. Wichtig dafür ist das Prinzip der graphischen Prägnanz. Mittels graphischer Kontraste kann inhaltlich Wichtiges auch wahrnehmungspsychologisch in den Vordergrund gerückt werden. Alles weniger Wichtige wird hingenommen in den Hintergrund verschoben.



Gestaltpsychologen haben Prinzipien postuliert, die visuelle Prägnanz schaffen, um bewusst Gestalt vom Hintergrund zu unterscheiden.



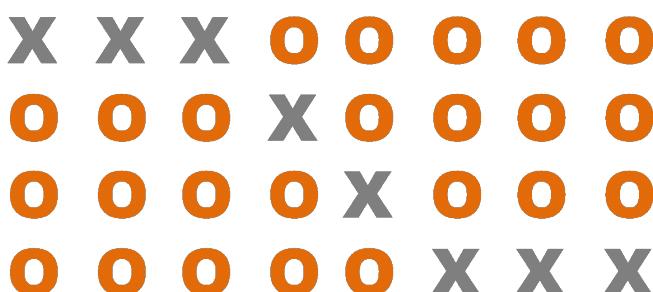
Um zweideutige Informationen zu verhindern, kann das Prinzip der Nähe angewandt werden. Elemente, die graphisch Gruppen bilden, werden als zusammengehörig identifiziert. Auch die Ähnlichkeit ist wichtig bei der Erkennung von zusammengehöriger Information.



Ein weiteres wichtiges Prinzip ist das der Geschlossenheit. Trotz der unterbrochenen Linien nehmen wir die Figur als oranges Quadrat wahr.



Auch das Prinzip der Kontinuität hilft, Gruppen zu erkennen.



4.5 Lektion 5: Wozu visualisiere ich?

Diese Lektion hat folgende Lernziele:

- Sie können gezielte Frage stellen um Mengen, Verteilungen, Anteile, oder Zusammenhänge zu visualisieren.
- Sie sind befähigt, den geeigneten Diagrammtyp für Ihre Fragestellung auszuwählen.
- Sie können diverse Visualisierungslösungen vorschlagen, um Mengen, Verteilungen, Anteile oder Zusammenhänge effektiv darzustellen.

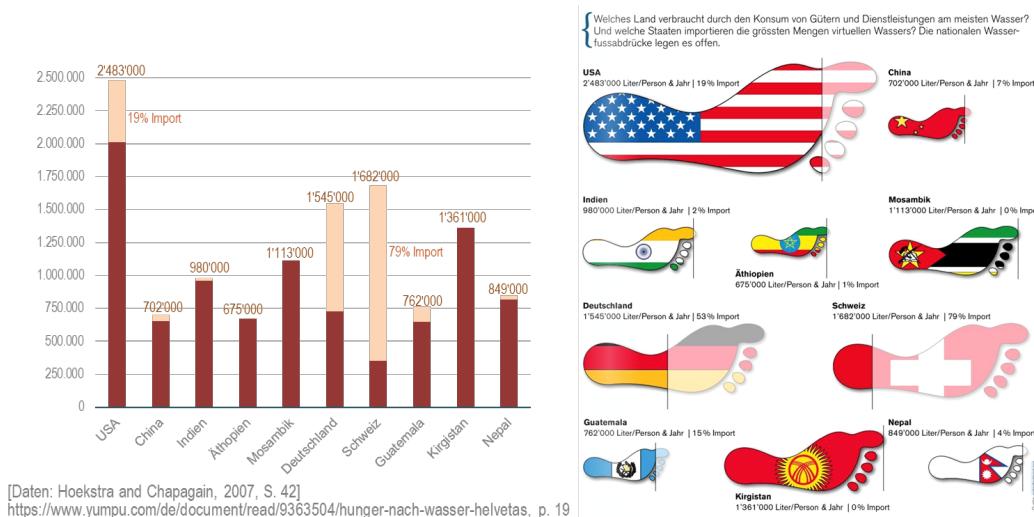
4.5.1 Visualisierung von Mengen

“Wie viel gibt es von x, y oder z in meinem Datensatz?” ist hier die Frage. Ein Beispiel dafür ist der durchschnittliche virtuelle Wasserfußabdruck von Ländern im globalen Vergleich). Als Grundlage für die Visualisierung dient die Tabelle 2.

Land	Verbrauch (Liter/Person und Jahr)	Import (%)
USA	2'483'000	19
China	702'000	7
Indien	980'000	2
Äthiopien	675'000	1
Mosambik	1'113'000	0
Deutschland	1'545'000	53
Schweiz	1'682'000	79
Guatemala	762'000	15
Kirgistan	1'361'000	0
Nepal	849'000	4

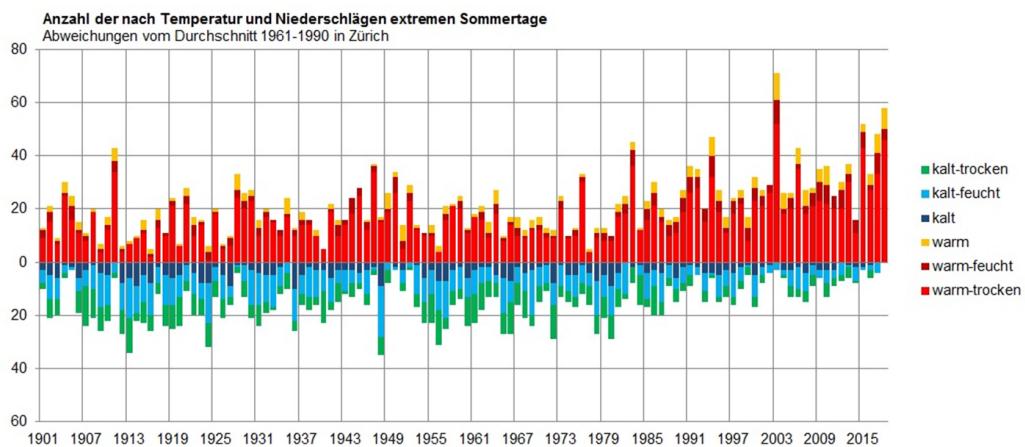
Tabelle 2: Datentabelle zum Wassermengenvergleich

Zur Visualisierung dieses Mengenvergleichs könnten gestaffelte Säulen- oder Balkengrafiken verwendet werden.



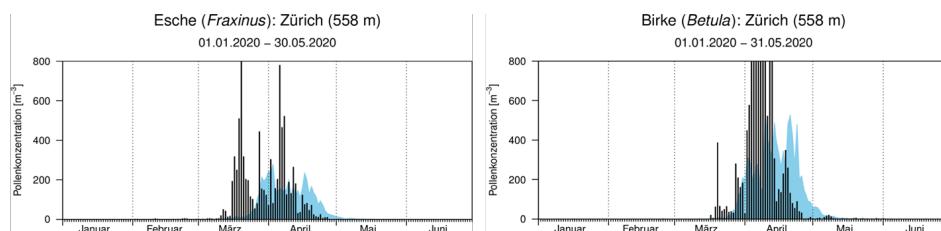
4.5.2 Visualisierung einer Verteilung

Eine Beispielfrage ist hier ob der Sommer heisser wird in der Stadt Zürich. Zur Visualisierung von Verteilungen könnte man mit Histogrammen arbeiten.



4.5.3 Vergleich mehrerer Verteilungen

“Wie stark war die Pollenbelastung von Eschen und Birken in Zürich während der Saison 2020 im Vergleich zum langjährigen Durchschnitt” ist hier die Frage. Zu ihrer Beantwortung wird wieder ein Histogramm verwendet, dieses Mal allerdings in Kombination mit Dichten.

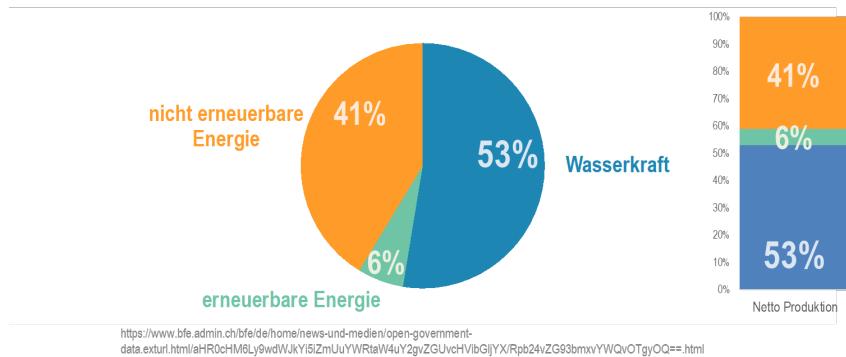


https://www.meteoschweiz.admin.ch/content/dam/meteoswiss/de/Ungebundene-Seiten/Publikationen/Klimabulletin/doc/klimabulletin_fruhling_2020_d.pdf

4.5.4 Visualisierung von Anteilen

Hier geht es um die Frage, wie gross die Anteile A, B und C an Total-X sind. Konkret soll dies zuerst an folgender Frage betrachtet werden: “Wie gross ist der Anteil der Wasserkraft an der gesamten Stromproduktion in der Schweiz?”. Zur Be-

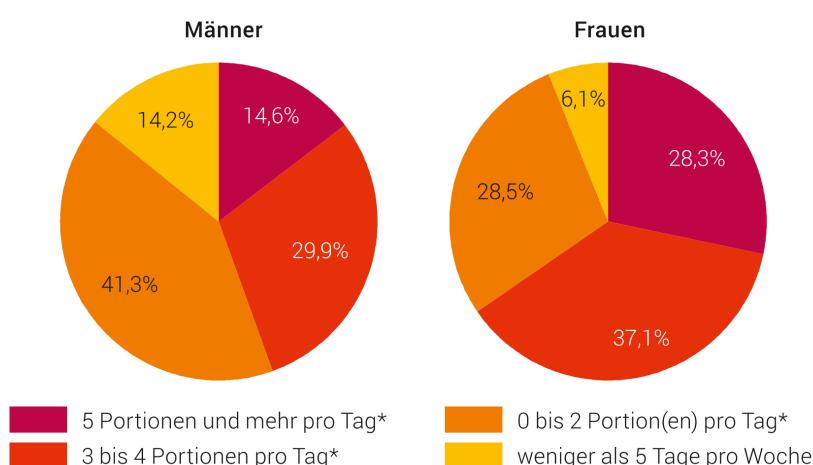
antwortung dieser Frage kann man sowohl Kreissektorendiagramme wie auch gestaffelte Säulen oder Balken verwenden.



Diese Diagrammtypen können aber auch zum Vergleich mehrerer Anteile verwendet werden, was anhand der Frage “Wie gross ist der Anteil der Personen in der Schweiz, die mindestens fünf Portionen Obst oder Gemüse pro Woche konsumieren? Gibt es Geschlechterunterschiede?” veranschaulicht werden soll.

Obst- und Gemüsekonsum, 2017

Bevölkerung ab 15 Jahren in Privathaushalten



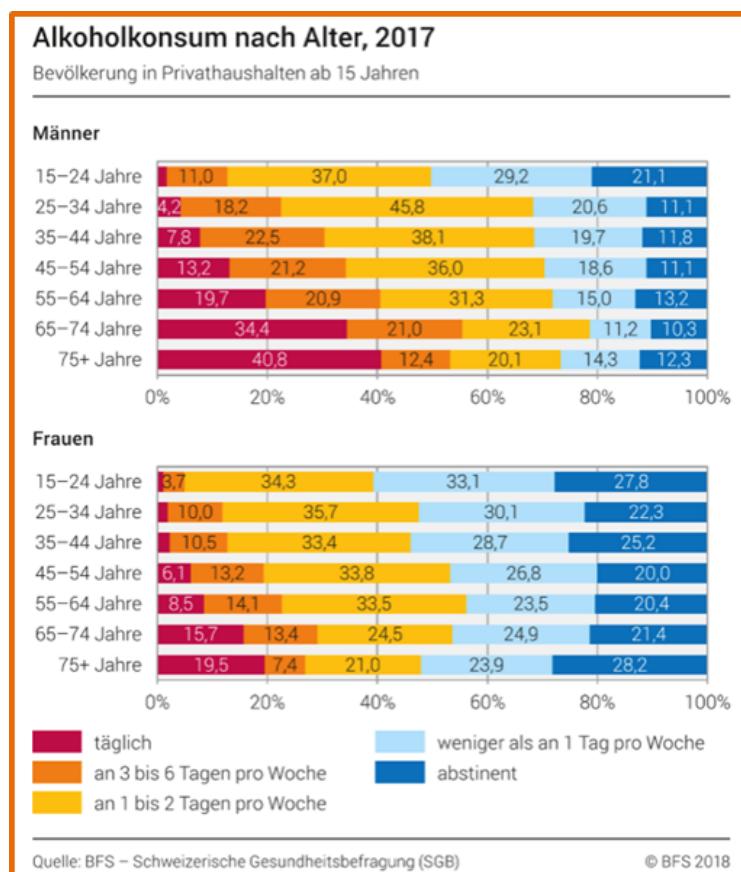
Quelle: BFS – Schweizerische Gesundheitsbefragung (SGB)

© BFS 2018

Beachtenswert ist hier, dass für grosse Anteile an Obst und Gemüse die Farbe Rot verwendet wurde. Diese wird in der westlichen Welt mit Gefahr assoziiert. Die Frage, ob die Entwickler dieser Grafik uns mitteilen wollten, dass viel Obst und Gemüse gesundheitsschädlich ist, bleibt offen.

Ein weiteres Beispiel für die Visualisierung mehrerer Anteile wird anhand der Fra-

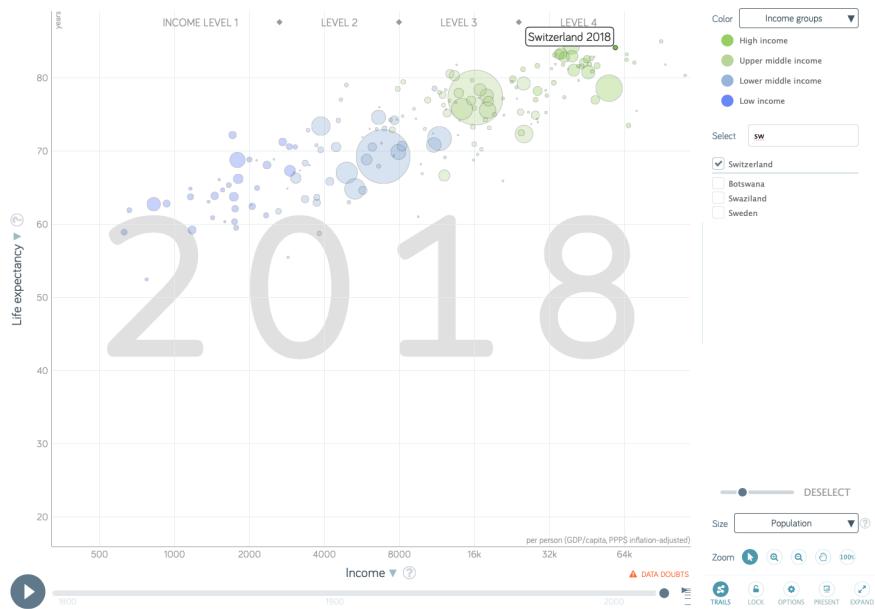
ge "Welche Altersklasse hat den höchsten Anteil am täglichen Alkoholkonsum in der Schweiz? Gibt es Geschlechterunterschiede?" gezeigt.



Hier ist die Farbwahl sehr gelungen. Kategorien mit häufigen Alkoholkonsum, der bekanntlich gesundheitsschädlich ist, sind in Rot-Gelb-Tönen dargestellt. Die gesünderen Kategorien hingegen in Blautönen, die an Wasser erinnern.

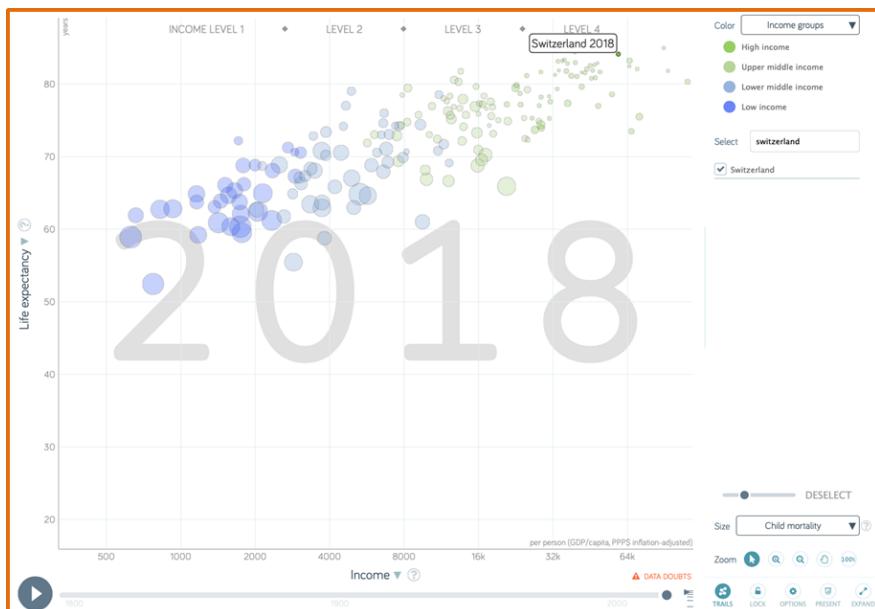
4.5.5 Visualisierung von Zusammenhängen

Hier geht es um die Frage, wie X und Y sich zueinander verhalten. Die einleitende Frage ist folgende: "Besteht ein Zusammenhang zwischen dem Einkommen und der Lebenserwartung in der Weltbevölkerung". Zur Beantwortung dieser Frage verwenden typischerweise Streudiagramme wie das untenstehende verwendet.



In diesem Koordinatensystem zeigt die y-Achse die Lebenserwartung der Bevölkerung eines Staates und die x-Achse stellt das Einkommen dar. Wir sehen sofort, dass reiche Länder eine höhere Lebenserwartung haben. Eine dritte Variable - die Menge der Bevölkerung pro Staat - ist durch die Grösse der Kreise abgebildet.

Eine weiteres Beispiel: “Besteht ein Zusammenhang zwischen dem Einkommen, der Lebenserwartung und der Kindersterblichkeit der Weltbevölkerung?”



<https://www.gapminder.org/>

In dieser Grafik wird die Grösse der Kreise durch die Kindersterblichkeit variiert. Aus dieser Grafik lässt sich erkennen, dass nicht nur die Lebenserwartung mit dem Wohlstand eines Staates steigt, sondern auch die Kindersterblichkeit sinkt.

4.6 Lektion 6: Wie visualisiere ich?

Die Lektion hat die folgenden Lernziele:

- Sie wissen welche Diagrammkomponenten bearbeitet werden sollten, um die Lesbarkeit der Visualisierung zu erhöhen.
- Sie kennen die graphischen Variablen, um Diagrammtypen attraktiv und gut lesbar zu gestalten.
- Sie können diverse graphische Lösungen vorschlagen, um Mengen, Verteilungen, Anteile oder Zusammenhänge prägnant darzustellen.

Für jeden Diagrammtyp existieren zahlreiche Möglichkeiten, die Grundform zu verändern. Sie sollten sich nie auf die Standardeinstellungen einer Grafik-Software verlassen, egal ob es sich dabei um R oder Excel handelt. Automatisch erstellte Grafiken müssen eigentlich immer nachbearbeitet werden.

4.6.1 Verbesserungsmöglichkeiten für Mengendiagramme

Im Verlauf des Kursbausteins wurden bereits zwei Alternativen zu Säulen gezeigt: Skalierter Text und skalierte Füsse. Die vertikale Säulen können aber auch horizontalen Balken geändert werden. Dies macht dann Sinn, wenn die Beschriftungen der Säulen zu viel Platz einnehmen. Bei den Balken fallen ausserdem die horizontale Längenvergleiche leichter. Zudem wurden die Zwischenräume der Balken verkleinert, um die Vergleiche zu vereinfachen. Drittens wurde das rote Farbschema der Säulen durch Blautönen in den Balken ersetzt, da Wasser eher mit blau als mit rot assoziiert wird. So wird auch der Kontrast zu den in Orange hervorgehobenen Werten erhöht. Die Reihenfolge der Länder wurde neu ausserdem nach der Länge der dunkelblauen Balken, die die Menge des nicht-importierten Wassers darstellen, sortiert.

Standardmäßig wird in vielen Statistikpaketen automatisch eine Legendenbox ausserhalb des Diagramms generiert. Bei beiden Grafiken wurde hier darauf

verzichtet, da es sich lohnt, die Information wann immer möglich in die Grafik zu integrieren. Ein wichtiges Element einer Grafik ist der Titel. Dieser sollte die Betrachtenden auf die wichtige Information im Diagramm aufmerksam machen oder prägnant die Antwort auf die gestellte Frage liefern.

4.6.2 Verbesserungsmöglichkeiten für Verteilungsdiagramme

Histogramme zeigen Häufigkeiten, also Daten, die in Klassen eingeteilt sind. Diese Klassen können einen Wert oder Wertebereich darstellen. Im Gegensatz zu den Säulendiagrammen sollten die Säulen bei kontinuierlichen Wertebereichen keine Abstände aufweisen. Alle Verbesserungsmöglichkeiten, die bereits für Säulen- und Balkendiagramme genannt wurden, gelten auch für Histogramme.

Im Beispiel der Grafiken zur Altersstruktur sollte der Titel durch den Folientitel “Die Schweizer Bevölkerung wird immer älter” ersetzt werden, die dieser die Kernaussage der Abbildung wiedergibt. Wichtig sind auch klare Kontraste zwischen dem Hintergrund und Vordergrund einer Visualisierung.

4.6.3 Verbesserungsmöglichkeiten für Anteildiagramme

Excel ermöglicht das Erstellen von 3D-Kreissektorendiagrammen. Darauf sollte unbedingt verzichtet werden, wenn die Daten nicht 3D sind. Ein Hauptkritikpunkt an Kreissektorendiagrammen im Allgemein ist, dass ihre Winkel weniger gut geschätzt werden können als die Längen von Säulen- oder Balkendiagrammen. Dies wird in 3D durch die perspektivische Verzerrung noch verstärkt. Außerdem sollten nicht mehr als 2-4 Sektoren in einem solchen Diagramm dargestellt werden, da zu kleine “Tortenstücke” schwierig zu lesen sind. Sektoren sollten immer bei 0° bzw. auf 12 Uhr beginnen und im Uhrzeigersinn gelesen werden können. Zudem gelten die bisher gemachten Verbesserungsvorschläge natürlich auch hier.

4.6.4 Verbesserungsmöglichkeiten für Diagramme, die Zusammenhänge zeigen

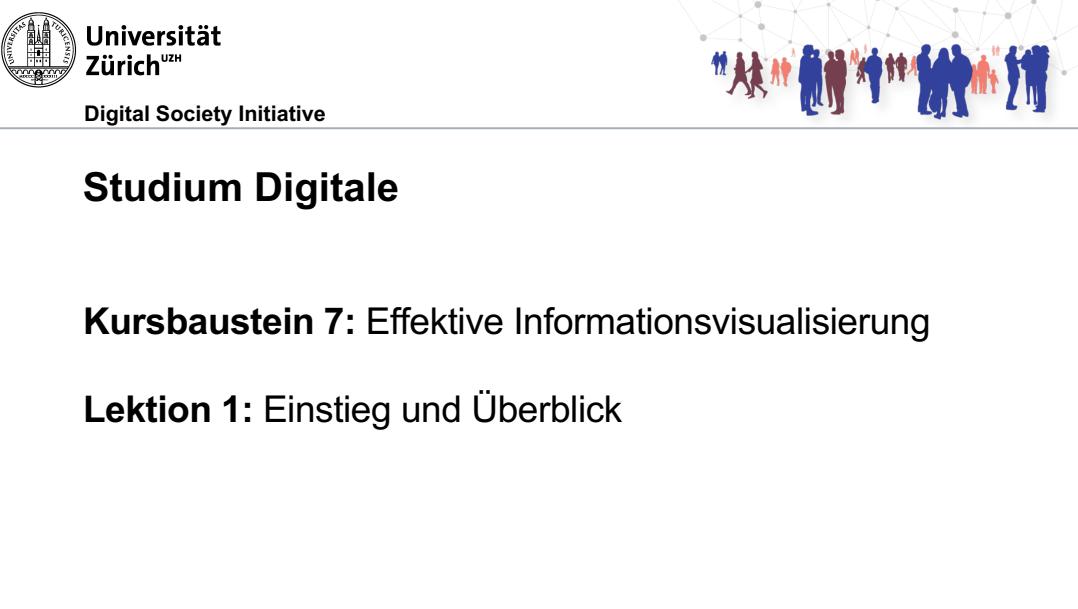
Vergleicht man die beiden Streudiagramme fällt auf, dass der Titel geändert, wichtige historische Daten hinzugefügt, sowie die Grösse und Farbe der Datenpunkte und der Schrift geändert wurden. Auch das Layout der Achsen wurde verändert und das Koordinatengitter entfernt. Mit dem vertikalen Strich können letztlich nun auch positive und negativen Werte besser unterschieden werden.

Referenzen

- Kanizsa, G., 1955. Margini quasi-percettivi in campi con stimulazione omogenea, Rivista di Psicologia, 49 (1). 7-30.

© Digital Society Initiative

4.7 Folien



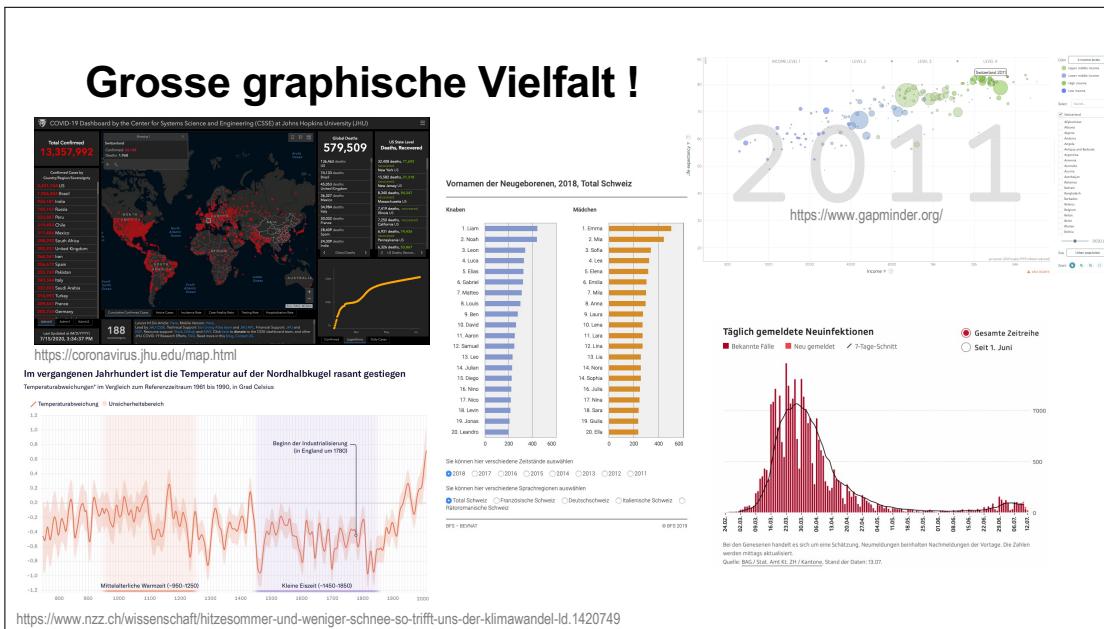
The banner features the University of Zurich logo (a circular seal with a building and text) next to the text "Universität Zürich UZH". Below this, the text "Digital Society Initiative" is displayed. To the right of the text is a graphic of stylized human figures in blue and red, connected by a network of grey dots, symbolizing connectivity and digital society.

Studium Digitale

Kursbaustein 7: Effektive Informationsvisualisierung

Lektion 1: Einstieg und Überblick





Lernziele

- Sie erkennen, wann es sich lohnt, ihre Zeit in eine Visualisierung zu investieren
- Sie kennen diverse Visualisierungsmethoden und können diese zieltgerecht einsetzen.
- Sie können die für Ihre Aufgabe und Daten angemessene Visualisierungsmethode auswählen und anwenden
- Sie können selbständig attraktive Visualisierungen mit Microsoft Excel erstellen



Überblick

1. Lernziele
2. Struktur: Die **5 Ws**

Die **5 Ws** der Visualisierung

Warum | Was | für Wen | Wozu und Wie wird visualisiert?



Warum visualisiere ich?

Erkenntnisgewinn aus
Datenhaufen

(07.12.17) Die Schweizer Wasserversorgungen stellten 2016 923 Mio. m³ Wasser bereit. Das sind gut 1% weniger als im Vorjahr. Der Pro-Kopf-Verbrauch sank von 307 auf 299 Liter pro Tag und liegt nach dem Trockenjahr 2015 nun wieder etwa auf dem Niveau von 2014. Damit setzt sich der rückläufige Verbrauchstrend fort, der in den Achtzigerjahren begann. Das geht aus der soeben fertiggestellten Wasserstatistik des Branchenverbands SVGW hervor. Eine Imageumfrage des Verbands ergab zudem, dass die Bevölkerung das Schweizer Trinkwasser weiterhin als sehr gut einstuft.

Bewegte sich in den 70er-Jahren der mittlere Tagesverbrauch pro Person noch an der 500-Liter-Marke, sank er seither stetig auf heute rund 300 Liter pro Tag. Auch die gesamte Wasserabgabe sank kontinuierlich und lag 2016 bei 923 Mio. m³, 1% weniger als 2015. Das entspricht rund dem 1.5-fachen Volumen des Murtensees. Hauptursache des markanten Mehrverbrauchs 2015 dürfte die ausgeprägte Trockenphase während der ganzen zweiten Jahreshälfte 2015 gewesen sein. Dafür spricht auch die gegenüber 2014 und 2016 um fast 20% höhere maximale Tagesabgabe.

Kosten seit den Achtzigerjahren um 7 Prozent gestiegen

Aus der neuen Statistik geht auch hervor, dass sich die Betriebs- und Kapitalkosten 2016 auf rund 1.6 Mrd. Fr. oder 188 Fr. pro Einwohner beliefen. Teuerungsbereinigt sind die Betriebs- und Kapitalkosten 7% höher als 1980. Die jährlichen Investitionen in den Wasserversorgungen betrugen 2016 925 Mio. Fr. oder 110 Fr. pro Einwohner. Teuerungsbereinigt werden damit pro Einwohner 7% mehr investiert als noch 1980. In den Wasserversorgungen arbeiteten 2016 rund 6500 Personen, deren Pensum rund 3100 Vollzeitstellen entspricht. Neben der Bereitstellung von Wasser produzierten die Schweizer Wasserversorger 2016 auch 186 GWh Strom, überwiegend mittels Trinkwasserturbinen. Damit deckte die Branche fast die Hälfte ihres Stromverbrauchs aus eigener Produktion.

Quelle:

http://www.svgw.ch/index.php?id=267&no_cache=1&tx_ttnews%5Btt_news%5D=1716&cHash=4b8c803c421031222b5901f65cccd5413/index.php?id=267&no_cache=1&tx_ttnews%5Btt_news%5D=1716&cHash=4b8c803c421031222b5901f65cccd5413



(07.12.17) Die Schweizer Wasserversorgungen stellten 2016 923 Mio. m³ Wasser bereit. Das sind gut 1% weniger als im Vorjahr. Der Pro-Kopf-Verbrauch sank von 307 auf 299 Liter pro Tag und liegt nach dem Trockenjahr 2015 nun wieder etwa auf dem Niveau von 2014. Damit setzt sich der rückläufige Verbrauchstrend fort, der in den Achtzigerjahren begann. Das geht aus der soeben fertiggestellten Wasserstatistik des Branchenverbands SVGW hervor. Eine Imageumfrage des Verbands ergab zudem, dass die Bevölkerung das Schweizer Trinkwasser weiterhin als sehr gut einstuft.

Bewegte sich in den 70er-Jahren der mittlere Tagesverbrauch pro Person noch an der 500-Liter-Marke, sank er seither stetig auf heute rund 300 Liter pro Tag. Auch die gesamte Wasserabgabe sank kontinuierlich und lag 2016 bei 923 Mio. m³, 1% weniger als 2015. Das entspricht rund dem 1.5-fachen Volumen des Murtensees. Hauptursache des markanten Mehrverbrauchs 2015 dürfte die ausgeprägte Trockenphase während der ganzen zweiten Jahreshälfte 2015 gewesen sein. Dafür spricht auch die gegenüber 2014 und 2016 um fast 20% höhere maximale Tagesabgabe.

Kosten seit den Achtzigerjahren um 7 Prozent gestiegen

Aus der neuen Statistik geht auch hervor, dass sich die Betriebs- und Kapitalkosten 2016 auf rund 1.6 Mrd. Fr. oder 188 Fr. pro Einwohner beliefen. Teuerungsbereinigt sind die Betriebs- und Kapitalkosten 7% höher als 1980. Die jährlichen Investitionen in den Wasserversorgungen betrugen 2016 925 Mio. Fr. oder 110 Fr. pro Einwohner. Teuerungsbereinigt werden damit pro Einwohner 7% mehr investiert als noch 1980. In den Wasserversorgungen arbeiteten 2016 rund 6500 Personen, deren Pensum rund 3100 Vollzeitstellen entspricht. Neben der Bereitstellung von Wasser produzierten die Schweizer Wasserversorger 2016 auch 186 GWh Strom, überwiegend mittels Trinkwasserturbinen. Damit deckte die Branche fast die Hälfte ihres Stromverbrauchs aus eigener Produktion.

Quelle:

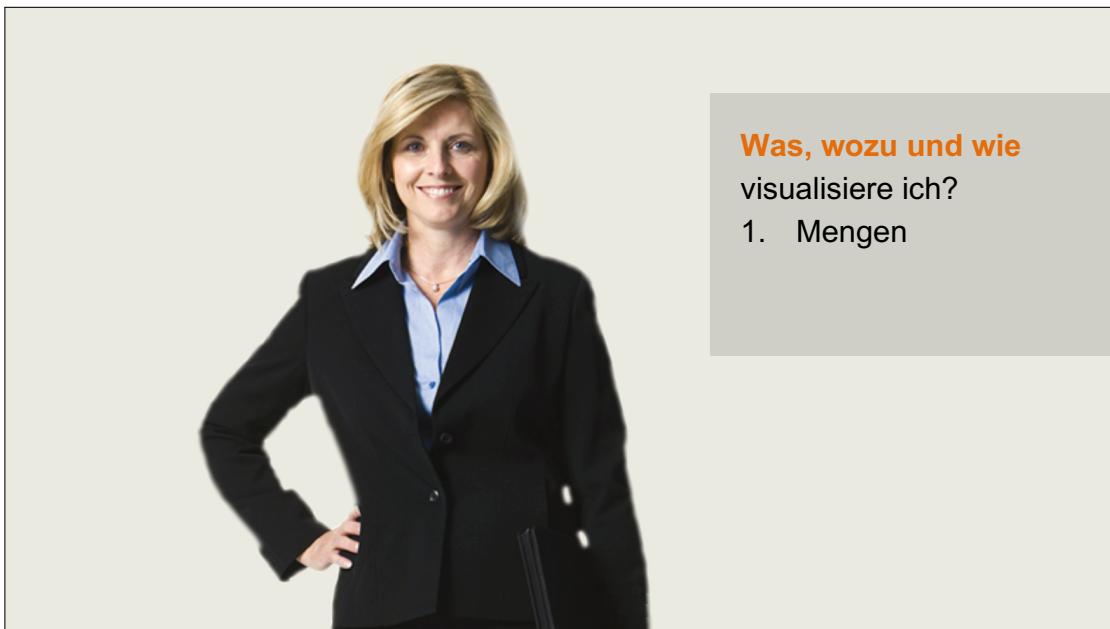
http://www.svgw.ch/index.php?id=267&no_cache=1&tx_ttnews%5Btt_news%5D=1716&cHash=4b8c803c421031222b5901f65cccd5413/index.php?id=267&no_cache=1&tx_ttnews%5Btt_news%5D=1716&cHash=4b8c803c421031222b5901f65cccd5413

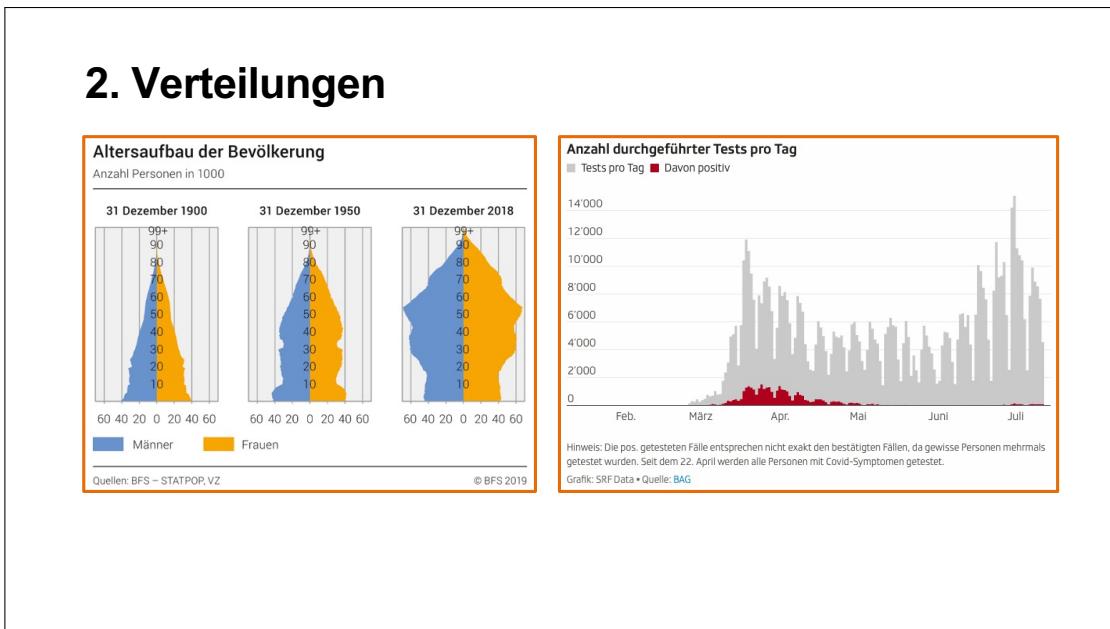


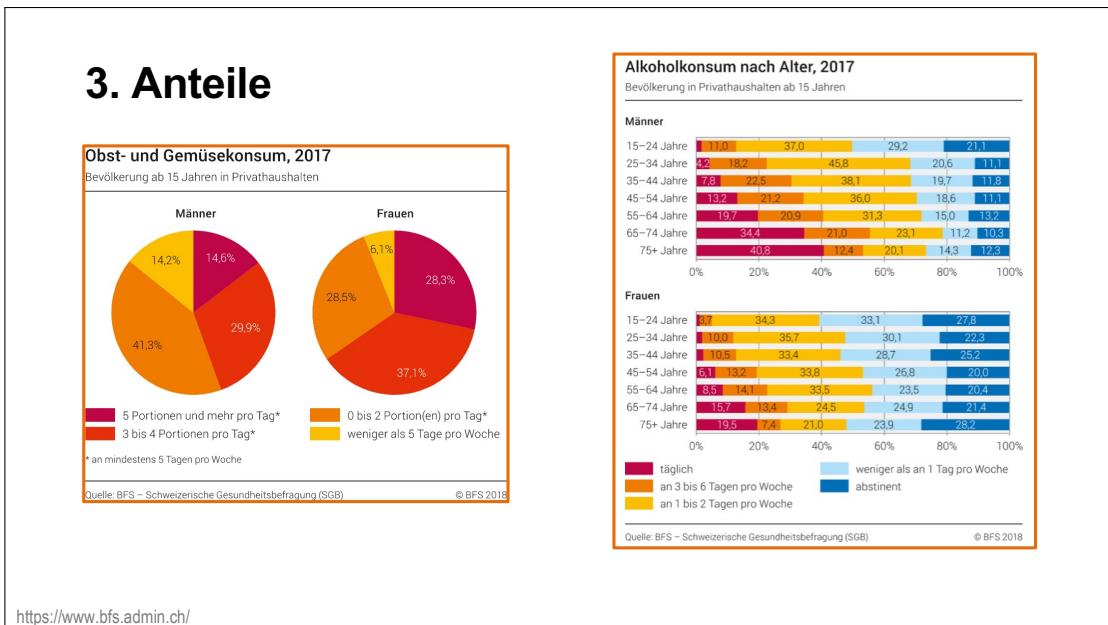
Für **wen** visualisiere ich?

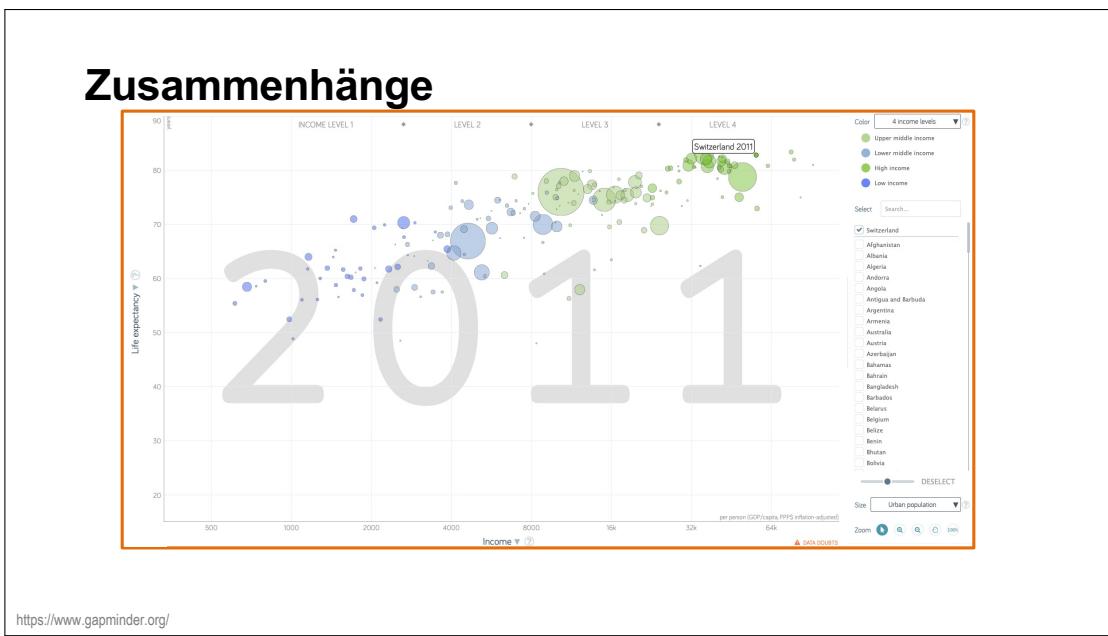
Wahrnehmungseffekte

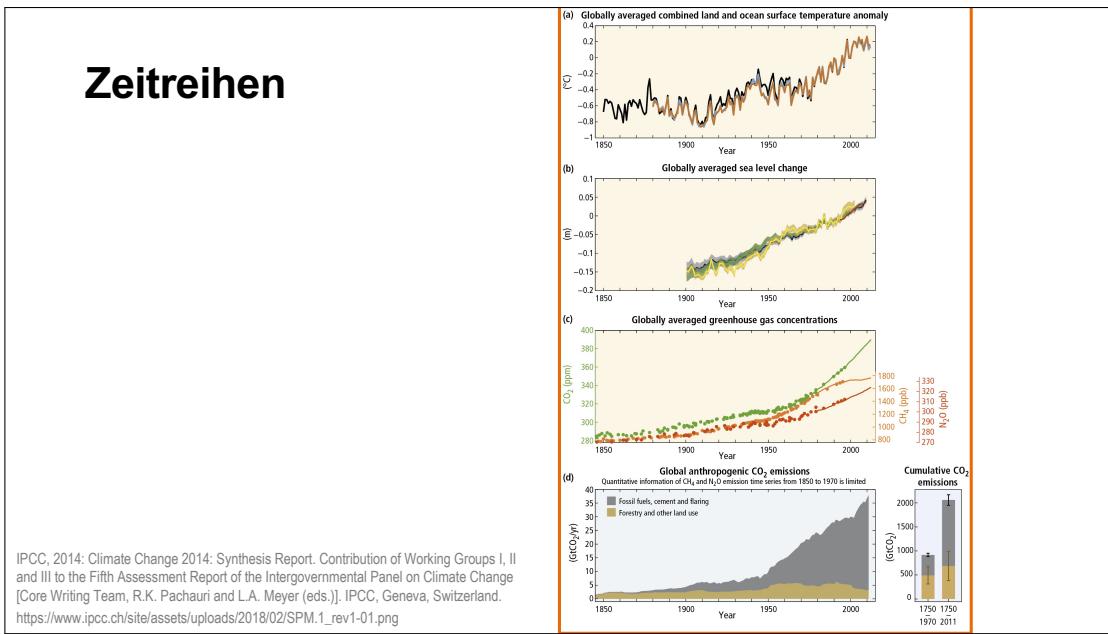
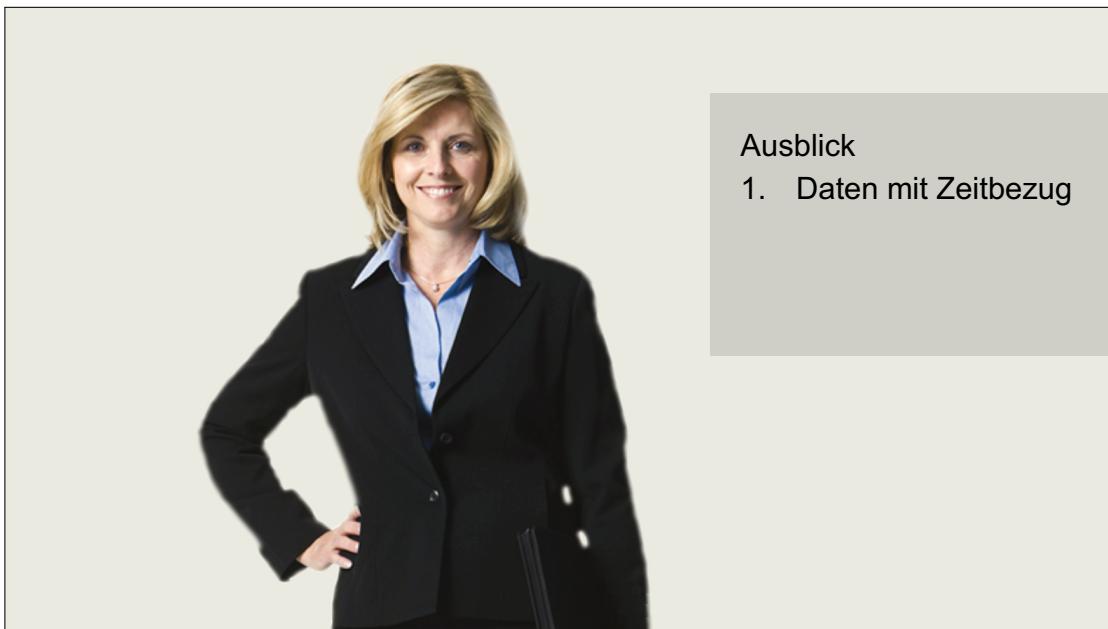


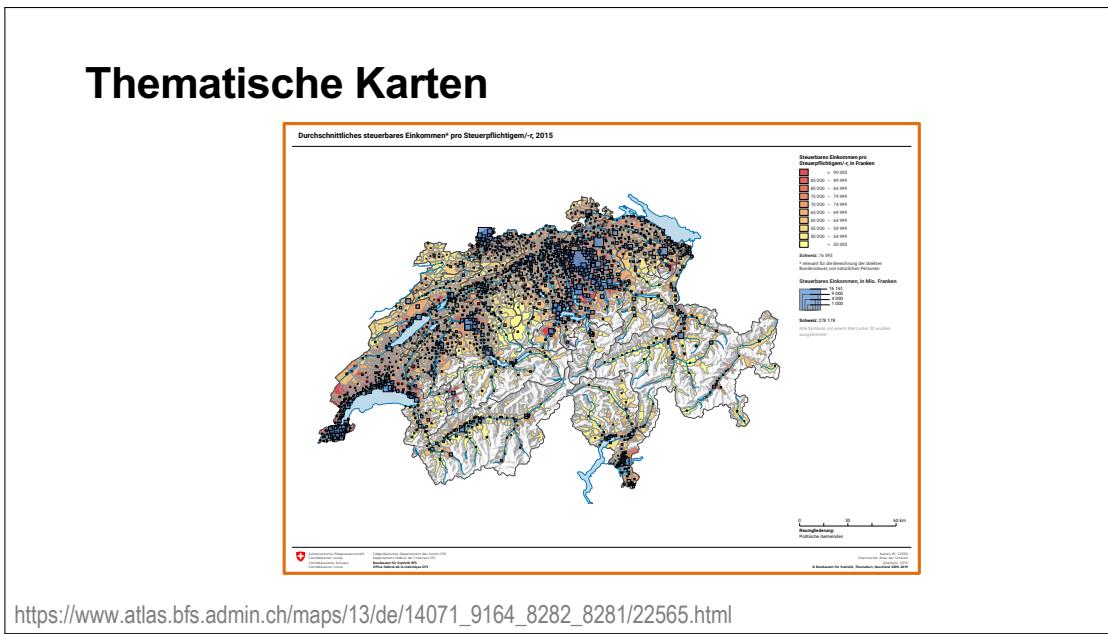
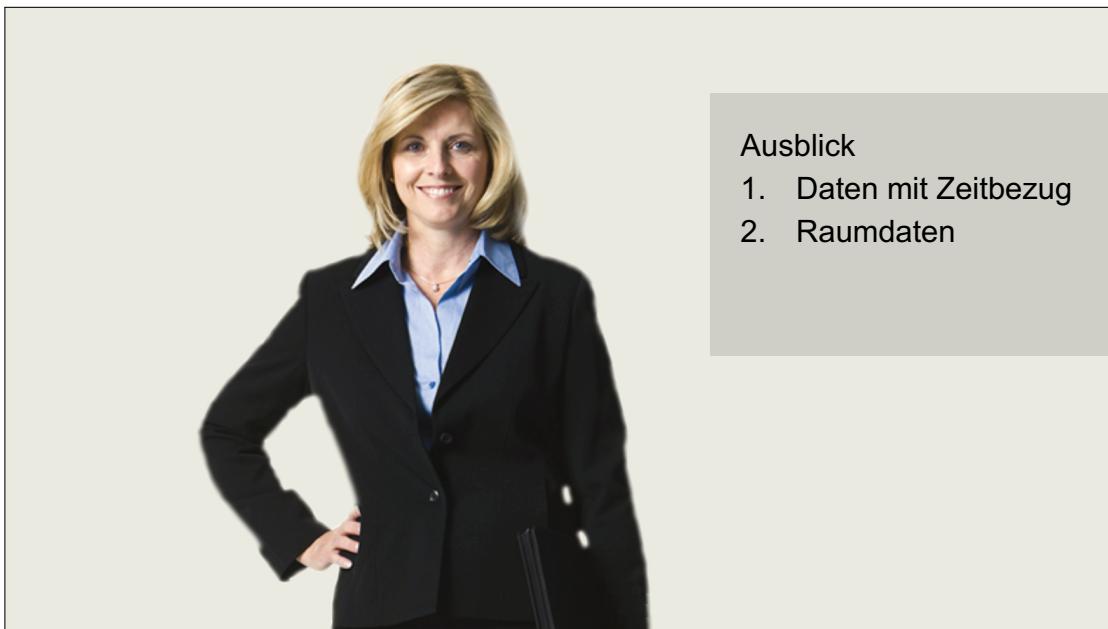












Hilfreiche Bücher über Datenvisualisierung

Claus Wilke

<https://serialmentor.com/dataviz/>

Stephanie Evergreen

<https://stephanieevergreen.com/books/>

Referenzen

Hilfreiche Bücher:

<https://serialmentor.com/dataviz/>

<https://stephanieevergreen.com/books/>

<https://www.amazon.com/gp/product/1492031089>

<https://www.amazon.de/Graphische-Semiolegric-Diagramme-Netze-Karten/dp/3110036606>

https://www.amazon.com/The-Elements-of-Graphing-Data/dp/0963488414/ref=cm_cr_dp_d_rvw_txt?ie=UTF8

<https://www.amazon.com/Visualization-Analysis-Design-AK-Peters/dp/1466508914>

<https://www.amazon.com/Design-Information-Introduction-Histories-Visualizations/dp/1592538061>

Weblinks

Hilfreiche Internetseiten:

<http://albertocairo.com>

<https://www.datawrapper.de>

http://www.visual-literacy.org/periodic_table/periodic_table.html

<https://flowingdata.com/2018/10/17/ask-the-question-visualize-the-answer/>

Referentin

Prof. Sara Irina Fabrikant

Geographisches Institut

© Universität Zürich

Digital Society Initiative



Universität
Zürich UZH

Digital Society Initiative



Studium Digitale

Kursbaustein 7: Effektive Informationsvisualisierung

Lektion 2: Warum will ich visualisieren?

Lernziele

- Sie können informiert entscheiden, wann es sich lohnt, Daten zu visualisieren.
- Sie können selbständig eine statistische Aussage mit einem Bild visualisieren.
- Sie können mit einem Beispiel illustrieren, warum die Visualisierung von Daten wichtig und zielführend ist.

Die 5 Ws in Visualisierung

Warum | Was | für Wen | Wozu und Wie wird visualisiert?



Warum visualisiere ich?

1. Kommunikation eines Sachverhalts

Der Wasserverbrauch in der Schweiz sinkt weiter

(07.12.17) Die Schweizer Wasserversorgungen stellten 2016 923 Mio. m³ Wasser bereit. Das sind gut 1% weniger als im Vorjahr. Der Pro-Kopf-Verbrauch sank von 307 auf 299 Liter pro Tag und liegt nach dem Trockenjahr 2015 nun wieder etwa auf dem Niveau von 2014. Damit setzt sich der rückläufige Verbrauchstrend fort, der in den Achtzigerjahren begann. Das geht aus der soeben fertiggestellten Wasserstatistik des Branchenverbands SVGW hervor. Eine Imageumfrage des Verbands ergab zudem, dass die Bevölkerung das Schweizer Trinkwasser weiterhin als sehr gut einstuft.

Bewegte sich in den 70er-Jahren der mittlere Tagesverbrauch pro Person noch an der 500-Liter-Marke, sank er seither stetig auf heute rund 300 Liter pro Tag. Auch die gesamte Wasserabgabe sank kontinuierlich und lag 2016 bei 923 Mio. m³, 1% weniger als 2015. Das entspricht rund dem 1.5-fachen Volumen des Murtensees. Hauptursache des markanten Mehrverbrauchs 2015 dürfte die ausgeprägte Trockenphase während der ganzen zweiten Jahreshälfte 2015 gewesen sein. Dafür spricht auch die gegenüber 2014 und 2016 um fast 20% höhere maximale Tagesabgabe.

Kosten seit den Achtzigerjahren um 7 Prozent gestiegen

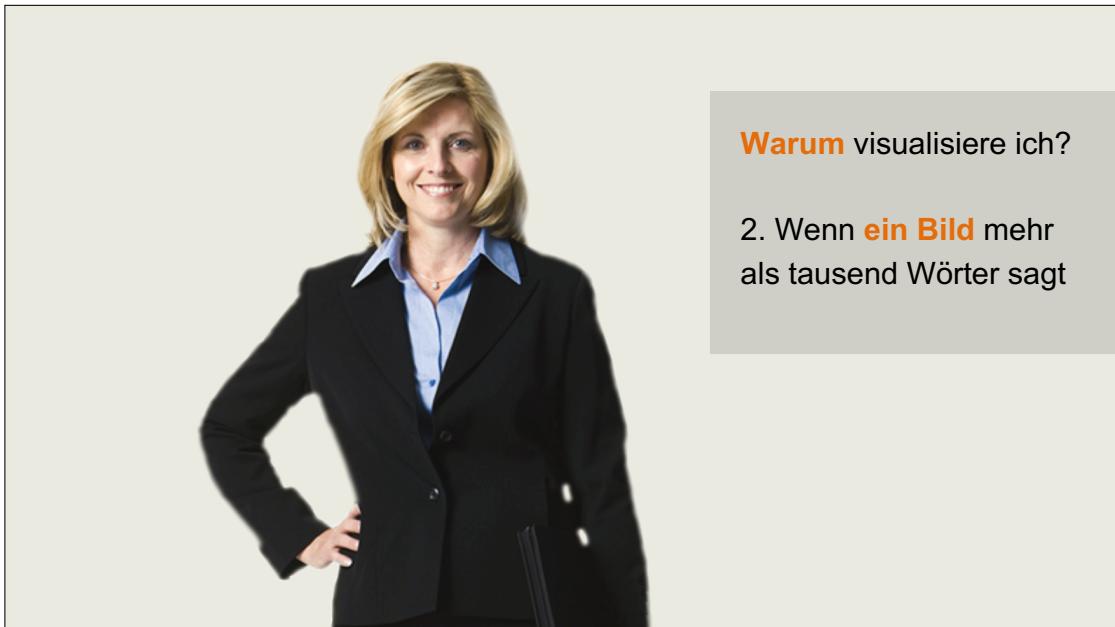
Aus der neuen Statistik geht auch hervor, dass sich die Betriebs- und Kapitalkosten 2016 auf rund 1.6 Mrd. Fr. oder 188 Fr. pro Einwohner beliefen. Teuerungsbereinigt sind die Betriebs- und Kapitalkosten 7% höher als 1980. Die jährlichen Investitionen in den Wasserversorgungen betrugen 2016 925 Mio. Fr. oder 110 Fr. pro Einwohner. Teuerungsbereinigt werden damit pro Einwohner 7% mehr investiert als noch 1980.

In den Wasserversorgungen arbeiteten 2016 rund 6500 Personen, deren Pensum rund 3100 Vollzeitstellen entspricht. Neben der Bereitstellung von Wasser produzierten die Schweizer Wasserversorger 2016 auch 186 GWh Strom, überwiegend mittels Trinkwasserturbinen. Damit deckte die Branche fast die Hälfte ihres Stromverbrauchs aus eigener Produktion.

Quelle: http://www.svgw.ch/index.php?id=267&no_cache=1&tx_ttnews%5Btt_news%5D=1716&cHash=4b8c803c421031222b5901f65cccd5413/index.php?id=267&no_cache=1&tx_ttnews%5Btt_news%5D=1716&cHash=4b8c803c421031222b5901f65cccd5413

Kommunikation eines Sachverhalts

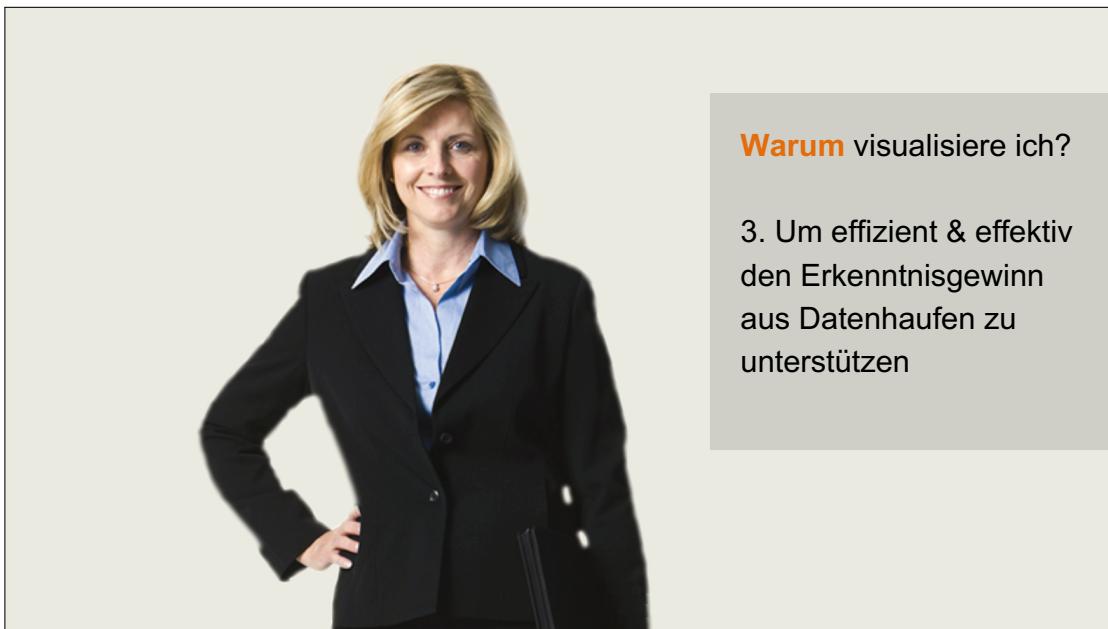
Wie hoch ist der mittlere Tagesverbrauch von Wasser pro Person in der Schweiz?



Warum visualisiere ich?

2. Wenn **ein Bild** mehr
als tausend Wörter sagt





Vergleich von vier überschaubaren Datensätzen

Set A		Set B		Set C		Set D	
X	Y	X	Y	X	Y	X	Y
10	8	10	9	10	7	8	7
8	7	8	8	8	7	8	6
13	8	13	9	13	13	8	8
9	9	9	9	9	7	8	9
11	8	11	9	11	8	8	8
14	10	14	8	14	9	8	7
6	7	6	6	6	6	8	5
4	4	4	3	4	5	19	13
12	11	12	9	12	8	8	6
7	5	7	7	7	6	8	8
5	6	5	5	5	6	8	7

[Anscombe 1973]

Anscombe, F. J. (1973). Graphs in Statistical Analysis. American Statistician. 27 (1): 17–21. doi:10.1080/00031305.1973.10478966.

Was steckt im Datenhaufen?

Welche Information bzw. Muster ist in den vier Datensätzen verborgen?

Vier Datensätze: identisches statistisches Muster !

Set A		Set B		Set C		Set D	
X	Y	X	Y	X	Y	X	Y
10	8	10	9	10	7	8	7
8	7	8	8	8	7	8	6
13	8	13	9	13	13	8	8
9	9	9	9	9	7	8	9
11	8	11	9	11	8	8	8
14	10	14	8	14	9	8	7
6	7	6	6	6	6	8	5
4	4	4	3	4	5	19	13
12	11	12	9	12	8	8	6
7	5	7	7	7	6	8	8
5	6	5	5	5	6	8	7

Mittelwert und Streuung

$$\mu_X = 9.0 \quad \sigma_X = 3.317$$

$$\mu_Y = 7.5 \quad \sigma_Y = 2.03$$

Linerarer Zusammenhang

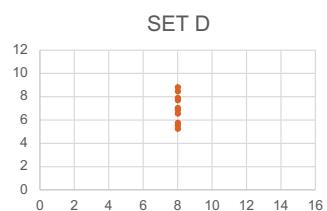
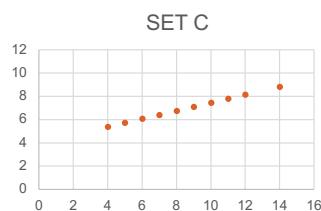
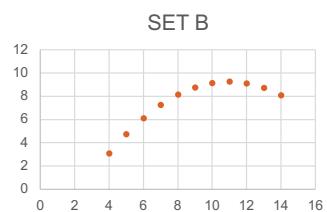
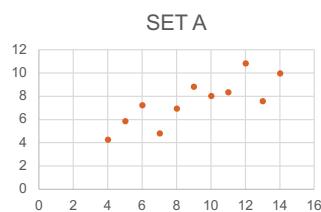
$$Y = 3 + 0.5X$$

$$R^2 = 0.67$$

[Anscombe 1973]

Anscombe, F. J. (1973). Graphs in Statistical Analysis. American Statistician. 27 (1): 17–21. doi:10.1080/00031305.1973.10478966.

Vier Datensätze: unterschiedliche graphische Muster !

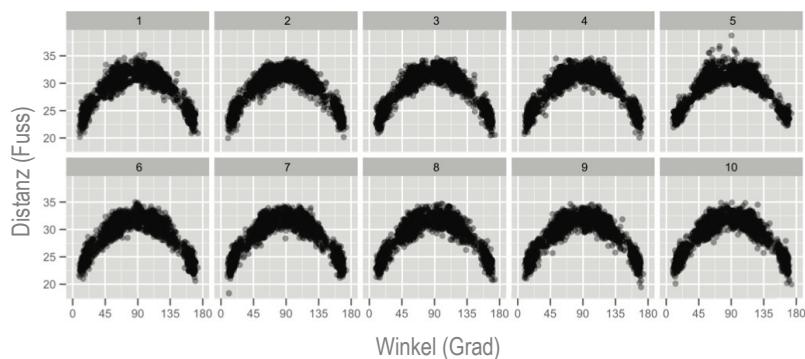


Warum visualisiere ich?

3. Um effizient & effektiv die Mustererkennung in Datenhaufen zu erleichtern.

Vergleich von 10 visualisierten Datensätzen:

Nullhypothese: Es besteht ein quadratischer Zusammenhang zwischen Wurfdistanz und Winkel zum Basketballkorb bei Drei-Punkte Würfen des Los Angeles Lakers Basketball Teams.

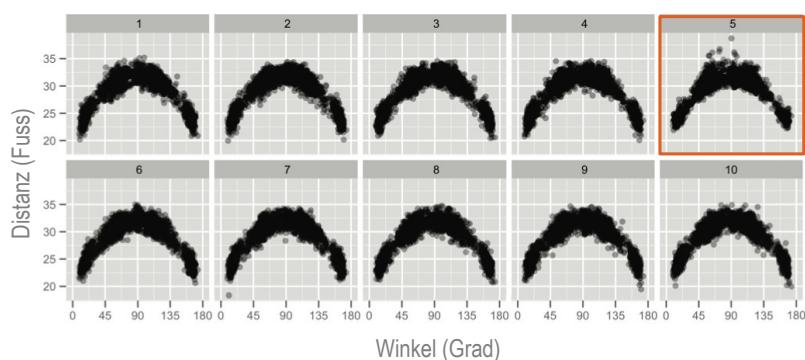


[Wickham et al. 2010]

Wickham, H., Cook, D., Hofmann, H., Buja, A. (2010). Graphical Inference for Infovis. IEEE Transactions on Visualization and Computer Graphics, vol. 16, no. 6: 973-979.

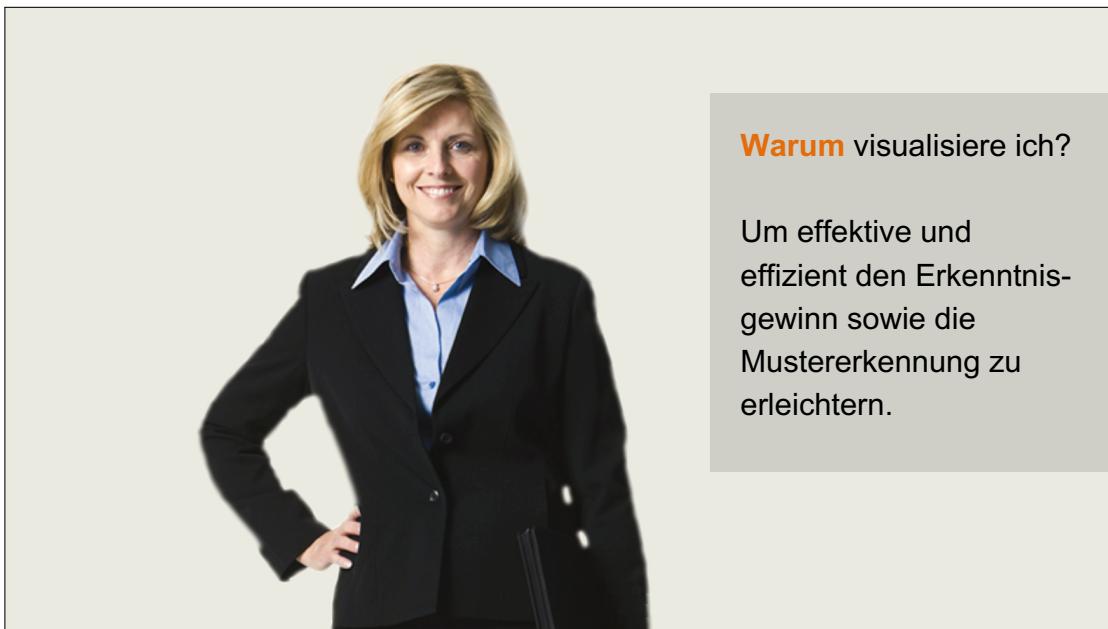
Vergleich von 10 visualisierten Datensätzen:

Nullhypothese: Es besteht ein quadratischer Zusammenhang zwischen Wurfdistanz und Winkel zum Basketballkorb bei Drei-Punkte Würfen des Los Angeles Lakers Basketball Teams.



[Wickham et al. 2010]

Wickham, H., Cook, D., Hofmann, H., Buja, A. (2010). Graphical Inference for Infovis. IEEE Transactions on Visualization and Computer Graphics, vol. 16, no. 6: 973-979.



Warum visualisiere ich?

Um effektive und effizient den Erkenntnisgewinn sowie die Mustererkennung zu erleichtern.

Referenzen

Anscombe, F. J. (1973). Graphs in Statistical Analysis. *American Statistician*. 27 (1): 17–21. doi:10.1080/00031305.1973.10478966.

Evergreen, Stephanie (2020). Effective Data Visualization. Second Edition, Sage, Thousand Oaks, (CA), U.S.A.

Wickham, H., Cook, D., Hofmann, H.. Buja, A. (2010). Graphical Inference for Infovis. *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 6: 973-979.

Wilke, Claus (2019). Fundamentals of Data Visualization. O'Reilly Media, Inc., Sebastopol, (CA), U.S.A.

Weblinks

<https://serialmentor.com/dataviz/>

<https://viz.wtf>

<http://www.svgw.ch>

<https://www.autodeskresearch.com/publications/samestats>

Referentin

Prof. Sara Irina Fabrikant

Geographisches Institut

© Universität Zürich

Digital Society Initiative



Universität
Zürich UZH

Digital Society Initiative



Studium Digitale

Kursbaustein 7: Effektive Informationsvisualisierung

Lektion 3: Was will ich visualisieren?

Lernziele

- Sie kennen zentrale Dateneigenschaften (Datentyp, Datenmerkmale, Messniveau, etc.). Siehe dazu auch den Kursbaustein «Statistik».
- Sie kennen die graphischen Variablen für die Visualisierung
- Sie haben einen Überblick über das Visualisierungsinventar

Die 5 Ws in Visualisierung

Warum | Was | für Wen | Wozu und Wie wird visualisiert?



Was visualisiere ich ?

1. Datentyp

wrapped by an almost perfectly symmetrical
volcano. The ceaseless beat of minute waves
the rocks I sit on, make me increasingly
worried and ~~sad~~^{relaxed} I saw my dreams today...
Dying hills, covered with small yellow
flowers, cut off abruptly, as knife in
a sea. I am in awe. Looking off
to the distance... snow capped mts
to the sky, beautifully rolling and smoothly
formed clouds. Vaguely appearing like
monsters in ~~the~~ ^{the} night. The sun
is still out, and is beginning to set. The
snow on the Oromo Volcano turns orange, and
the sky on the opposite horizon. As
I walk back through the Plaza, I see the
bus must be the hang-out area for the
young people at night. At the same time it
is where the older generation takes its walks.
I take a deep breath, and continue walking.

<https://pixabay.com/illustrations/pay-digit-number-fill-count-mass-1036469/>



<https://pixabay.com/illustrations/pay-digit-number-fill-count-mass-1036469/>



Was visualisiere ich?

1. Datentyp
2. Datenmerkmale



Welche Daten kann ich visualisieren?

Typ	Merkmal	Wert	Messniveau	Beschreibung
Numerisch	kontinuierlich	1.3, 10 ²	ratio	Distanz in Metern
Numerisch	diskret	1, 10, -25	interval	Temperatur in Grad Celsius
alphabetisch	diskret	niedrig, mittel, hoch	ordinal	Erfahrung in Visualisierung
alphabetisch	diskret	Rot, Grün, Blau	nominal	Farbton
Alpha-numerisch	Datum/Zeit	1. August 1291	Interval/ratio	Nationalfeiertag

Siehe Kurseinheit: "Statistik"

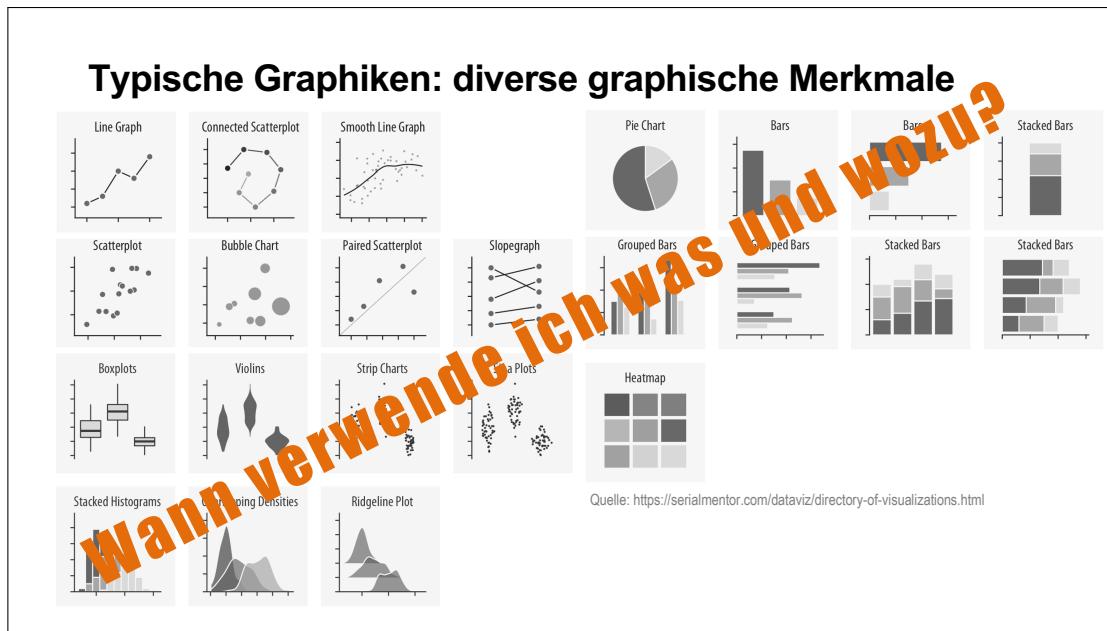
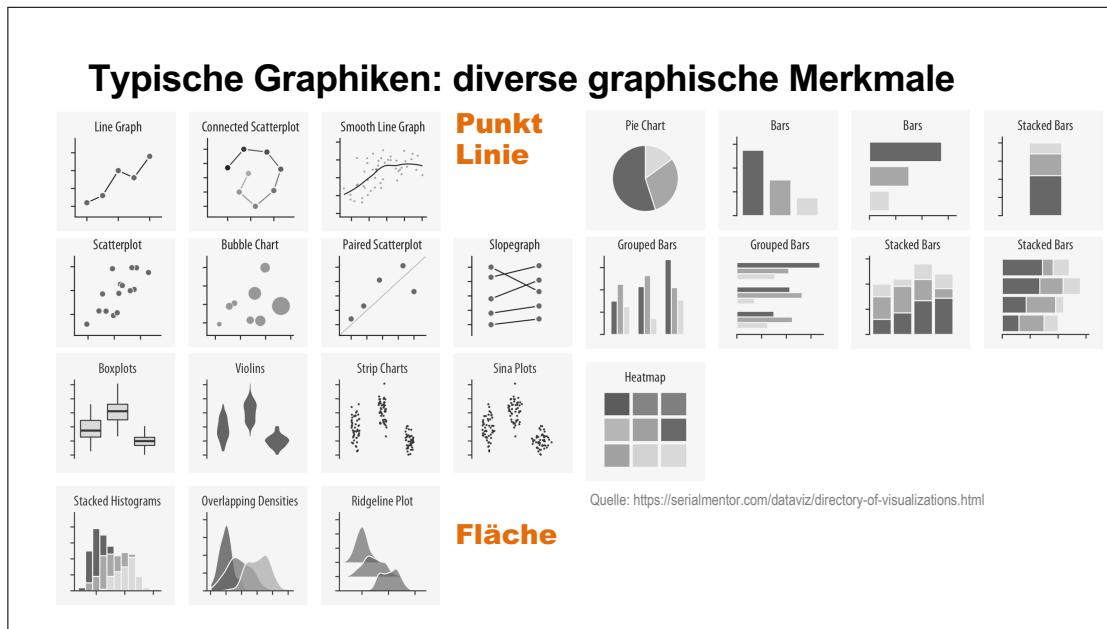
Eine typische Datentabelle

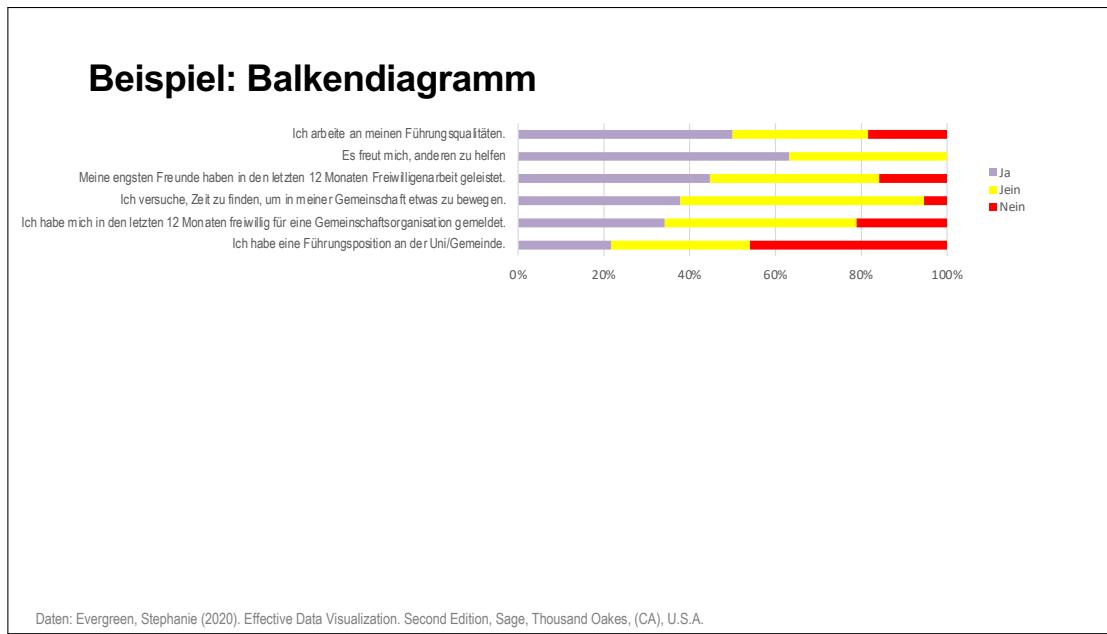
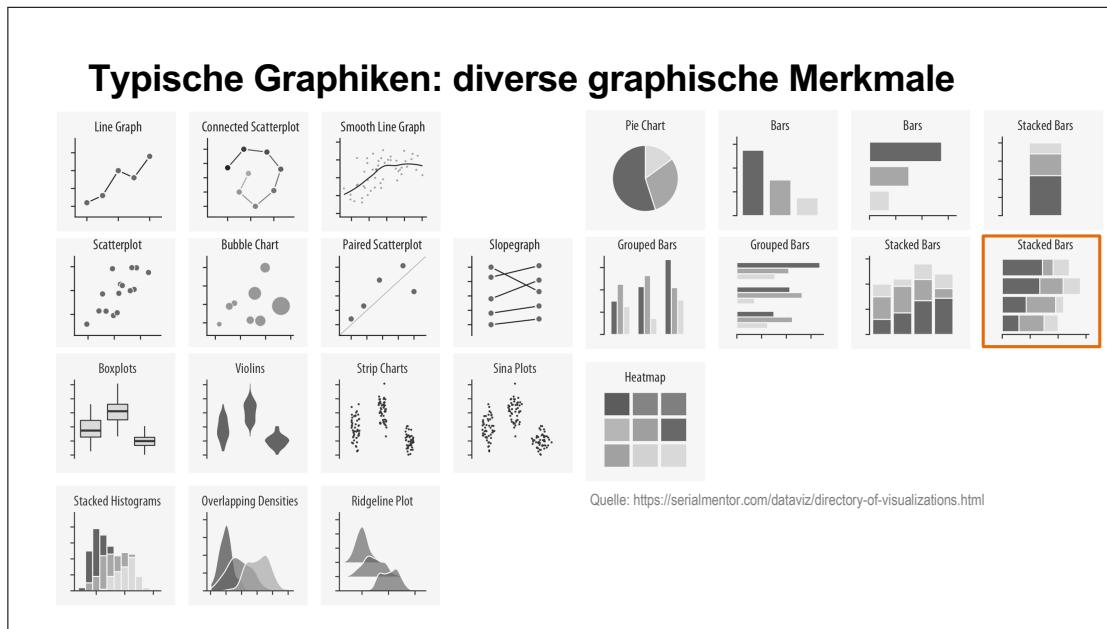
Jahr	Alterskategorie	Geschlecht	Kreis	Aufenthaltsdauer
1993	0-19	F	1	1.7
1993	0-19	F	2	5.1
1993	0-19	F	3	3.9
1993	0-19	F	4	3.4
1993	0-19	F	5	3.5

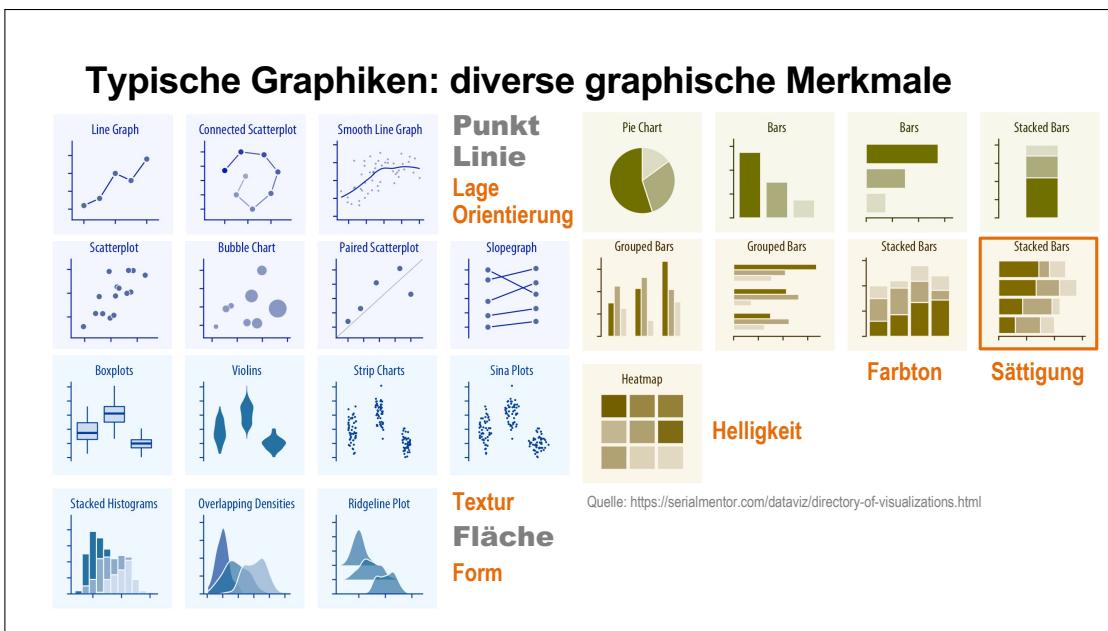
Quelle: <https://www.stadt-zuerich.ch/prd/de/index/statistik/themen/bevoelkerung/aufenthaltsdauer/aufenthaltsdauer-beim-wegzug.html>



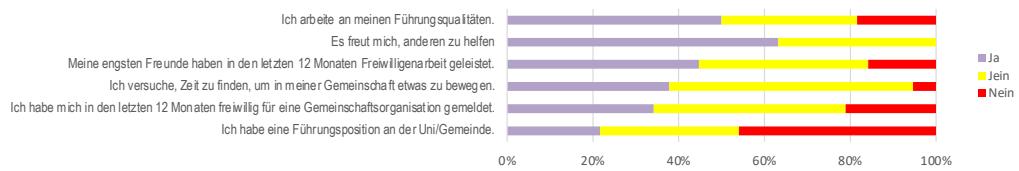
Was visualisiere ich?
1. Geometrie



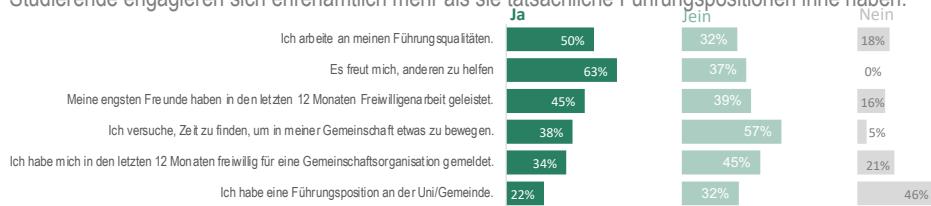




Gleicher Datensatz: unterschiedliche Darstellungen



Es besteht eine Lücke zwischen dem Erkennen von Führungschancen und dem Ergreifen dieser Chancen.
Studierende engagieren sich ehrenamtlich mehr als sie tatsächliche Führungspositionen inne haben.



Referenzen

Evergreen, Stephanie (2020). Effective Data Visualization. Second Edition, Sage, Thousand Oaks, (CA), U.S.A.

Wilke, Claus (2019). Fundamentals of Data Visualization. O'Reilly Media, Inc., Sebastopol, (CA), U.S.A.

Weblinks

<https://serialmentor.com/dataviz/>

<https://flowingdata.com/2018/10/17/ask-the-question-visualize-the-answer/>

Referentin

Prof. Sara Irina Fabrikant

Geographisches Institut

© Universität Zürich

Digital Society Initiative



Universität
Zürich UZH

Digital Society Initiative



Studium Digitale

Kursbaustein 7: Effektive Informationsvisualisierung

Lektion 4: Für **wen** wird visualisiert?

Lernziele

- Sie kennen typische visuelle Wahrnehmungseffekte
- Sie wissen um das Prinzip der graphischen Prägnanz
- Sie können die Bausteine des Gesetzes der guten Gestalt nennen
- Sie können die Gestaltbausteine selbstständig in Visualisierungen anwenden

Die 5 Ws in Visualisierung

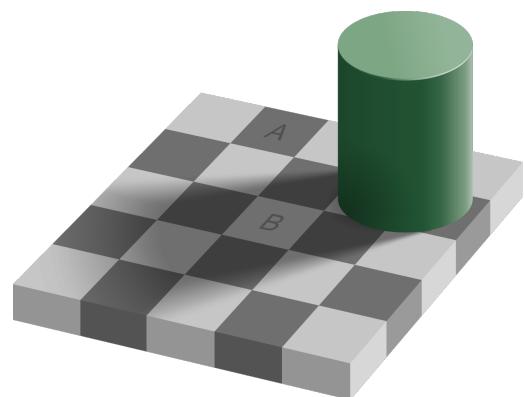
Warum | Was | für Wen | Wozu und Wie wird visualisiert?



Für **Wen** visualisiere ich?

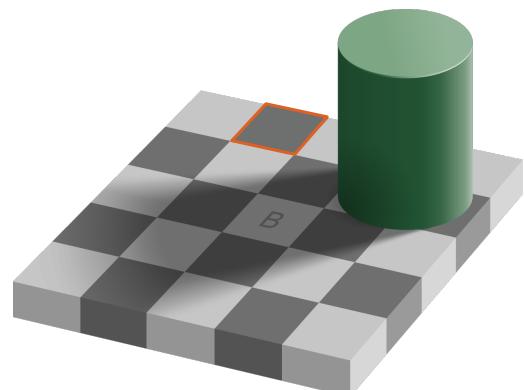
Für Personen mit einem Informationsbedürfnis unter Berücksichtigung deren visuellen Wahrnehmung.

Visuelle Realität...



Edward H. Adelson (1995)
<http://persci.mit.edu/gallery/checkershadow>

Visuelle Realität und Illusion!



Edward H. Adelson (1995)
<http://persci.mit.edu/gallery/checkershadow>



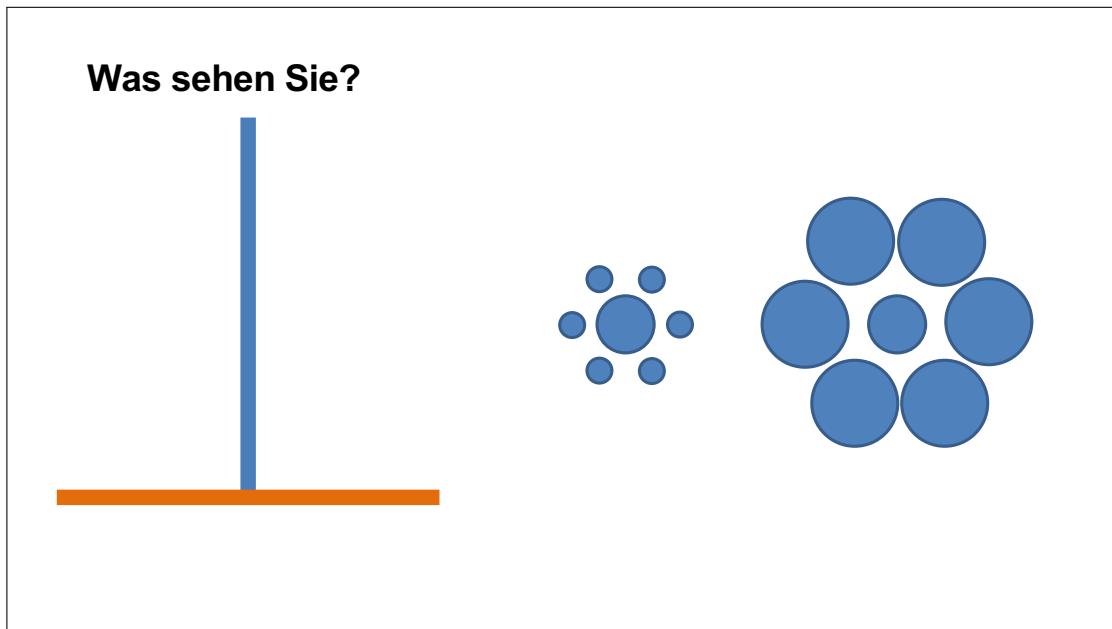
Menschliche Wahrnehmung
1. Perzeption

Was sehen Sie?

A diagram illustrating the Opel illusion. It consists of a short blue vertical line on a horizontal orange line. The blue line appears longer than the orange line due to the visual effect.

[Oppel 1815–1894]

<https://www.illusionsindex.org/ir/vertical-horizontal-illusion>



Was sehen Sie?

The slide features four separate color palettes arranged in a 2x2 grid. Each palette consists of a 4x6 grid of color swatches, with each swatch labeled with a name. The palettes are:

- Grayscale:** Burg, Orange, RedOr, Orange, DarkMint, Mint, Indigo, Sunset.
- Dark:** Hush, Haze, Indigo, Teal, DarkMint, Mint, Indigo, Sunset.
- Green:** Burg, Orange, RedOr, Orange, DarkMint, Mint, Indigo, Sunset, Teal, Teal.
- Yellow:** Burg, Orange, RedOr, Orange, DarkMint, Mint, Indigo, Sunset, Teal, Teal.

Farblegenden: <https://carto.com/cartocolors/>



Menschliche Wahrnehmung

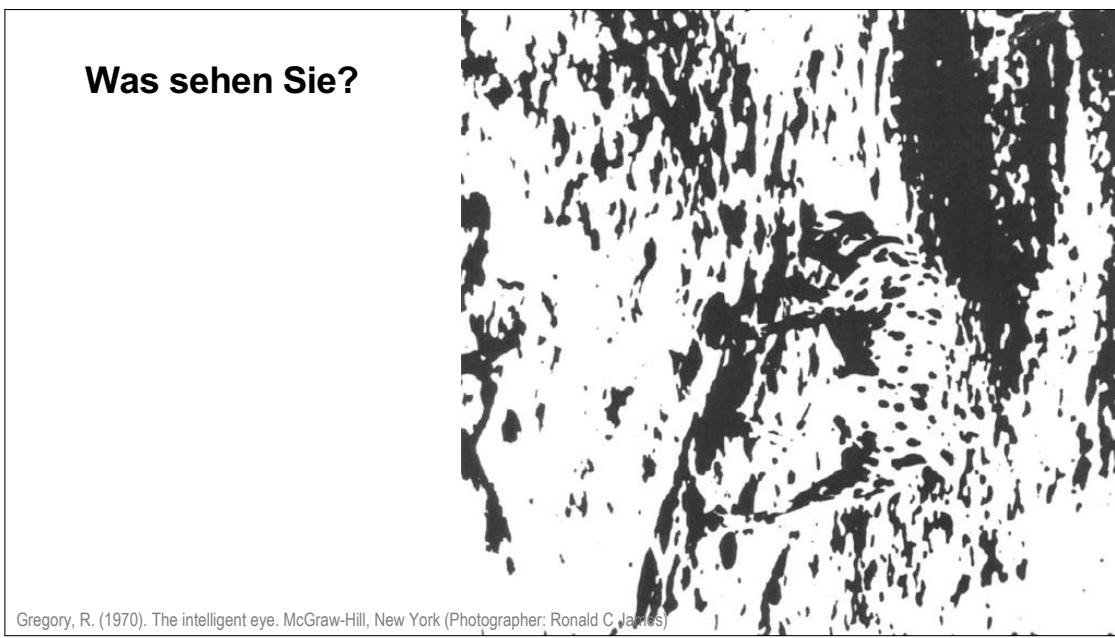
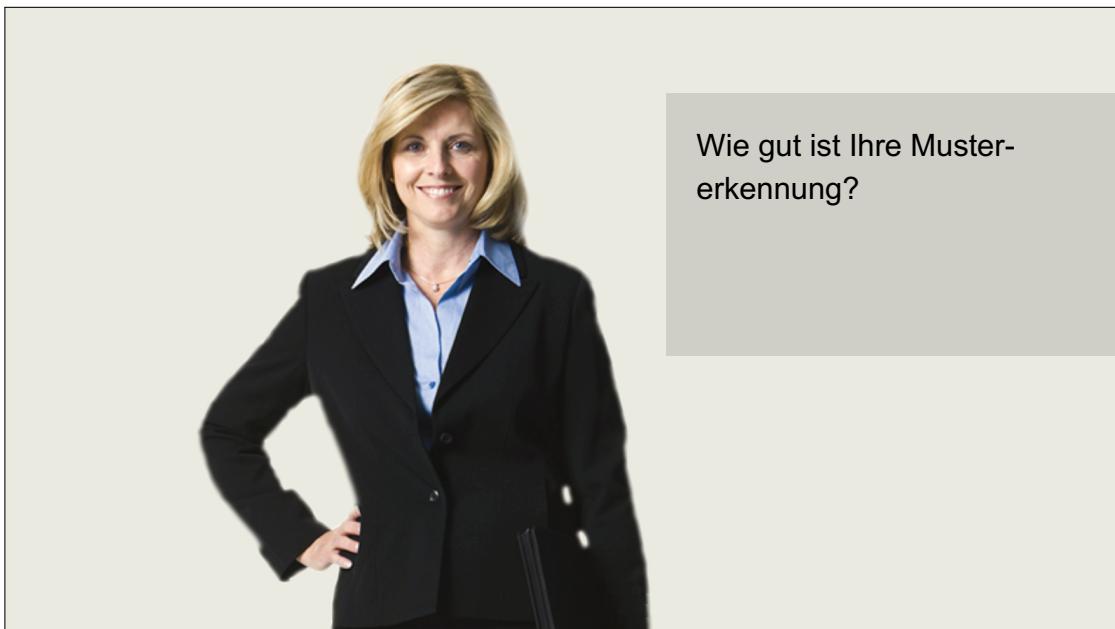
1. Perzeption
2. Kognition

Lesen Sie den Text laut vor...



The text sequence consists of ten words arranged in two columns. The first column contains five words: ROT, GRÜN, BLAU, ROT, GELB. The second column contains five words: GRÜN, BLAU, ROT, BLAU, GELB. Each word is in a different color: ROT is blue, GRÜN is yellow, BLAU is dark blue, and GELB is green.

Stroop, R. (1935) Studies of interference in serial verbal reactions. In: Journal of Experimental Psychology. Band 18, 1935, S. 643–662.



Was sehen Sie jetzt?



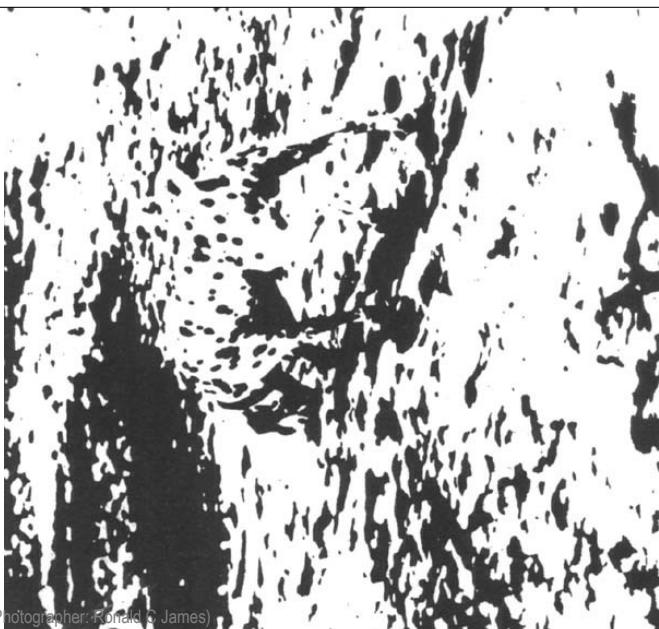
Gregory, R. (1970). The intelligent eye. McGraw-Hill, New York (Photographer: Ronald C James)

Was sehen Sie nun?



Gregory, R. (1970). The intelligent eye. McGraw-Hill, New York
(Photographer: Ronald C James)

Was sehen Sie?



Gregory, R. (1970). The intelligent eye. McGraw-Hill, New York (Photographer: Ronald O James)

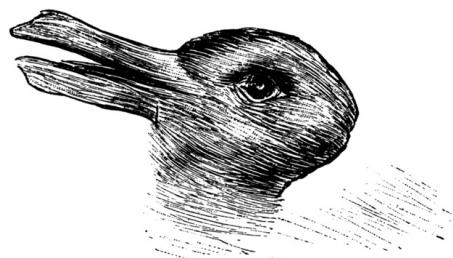
Was sehen Sie?



[anon. ca. 1888]

Was sehen Sie?

Welche Thiere gleichen ein-
ander am meisten?

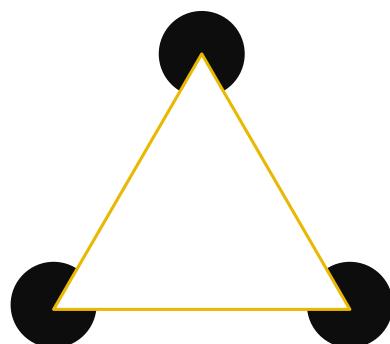


[Fliegende Blätter 23 October 1892]

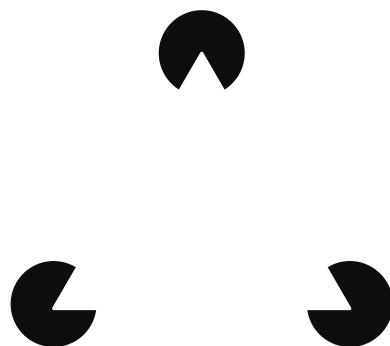
Was sehen Sie?



Ein weisses Dreieck?



Ein weisses Dreieck?



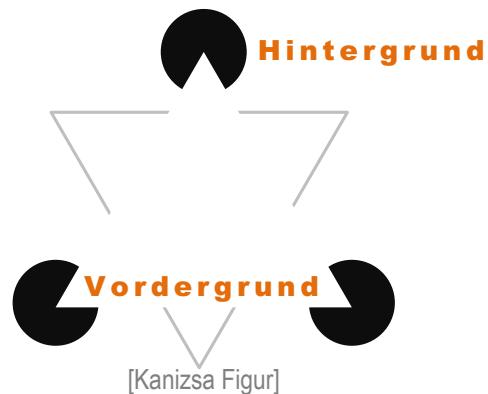


Was ist visuell prägnant?

The diagram illustrates the Kanizsa figure, a classic example of visual perception. It consists of several black and white shapes: two black circles at the top, two white triangles pointing downwards from behind them, and two black circles at the bottom. The text "Hintergrund" (Background) is placed next to the top circles, and "Vordergrund" (Foreground) is placed next to the bottom circles. A bracket below the bottom circles is labeled "[Kanizsa Figur]".

Kanizsa, G., 1955. Margini quasi-percettivi in campi con stimolazione omogenea, Rivista di Psicologia, 49 (1): 7–30.

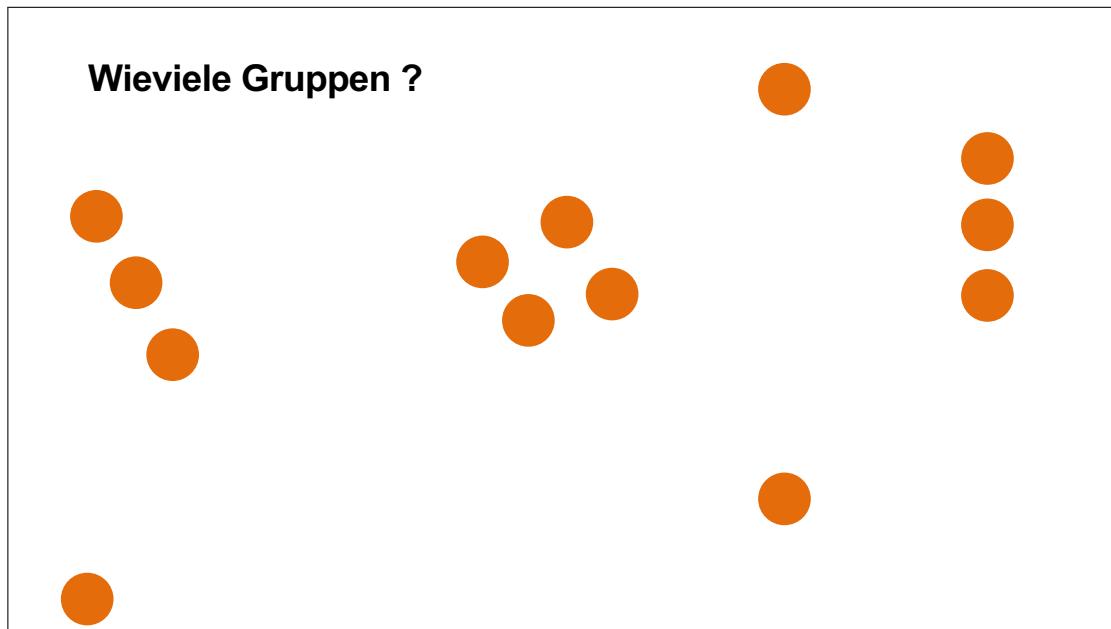
Was ist visuell prägnant?



Kanizsa, G., 1955. Margini quasi-percettivi in campi con stimolazione omogenea, Rivista di Psicologia, 49 (1): 7–30.

Wie erreiche ich visuelle Prägnanz?



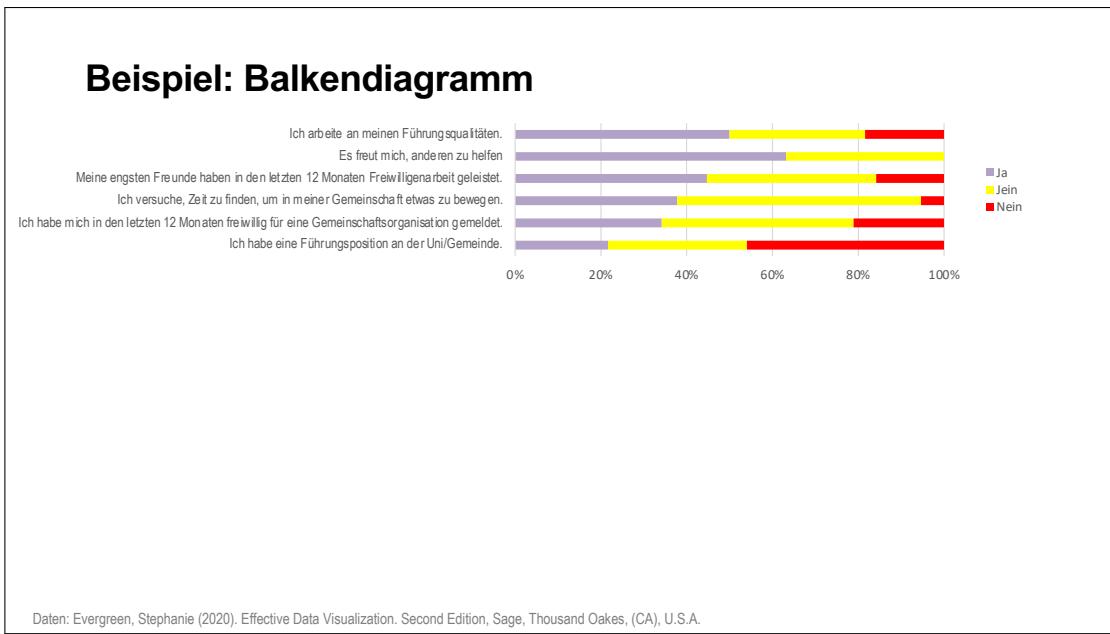
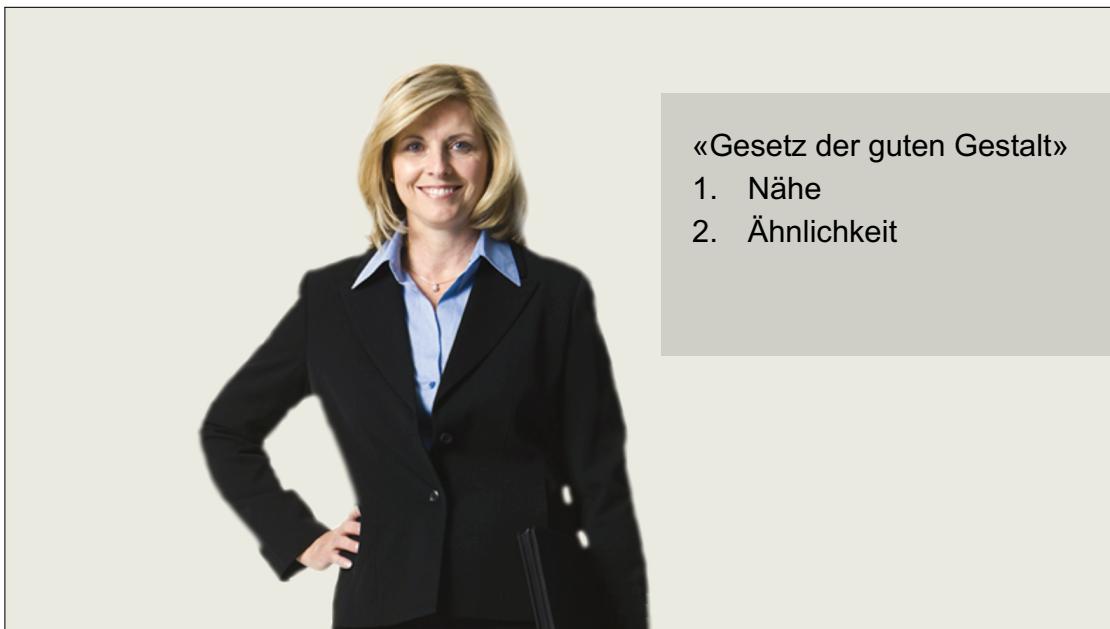


Wieviele Gruppen ?

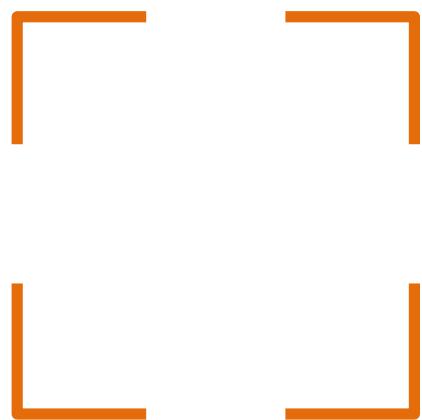
O X O X O X O X
O X O X O X O X
O X O X O X O X
O X O X O X O X

Wieviele Gruppen ?

O X O X O X O X
O X O X O X O X
O X O X O X O X
O X O X O X O X



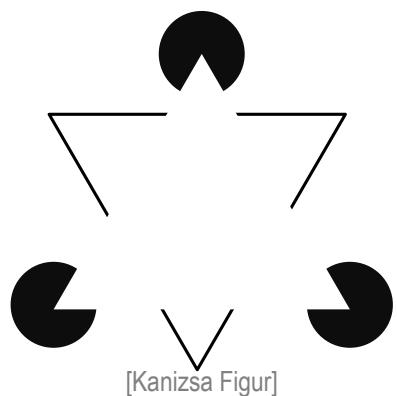
Wieviele Figuren?



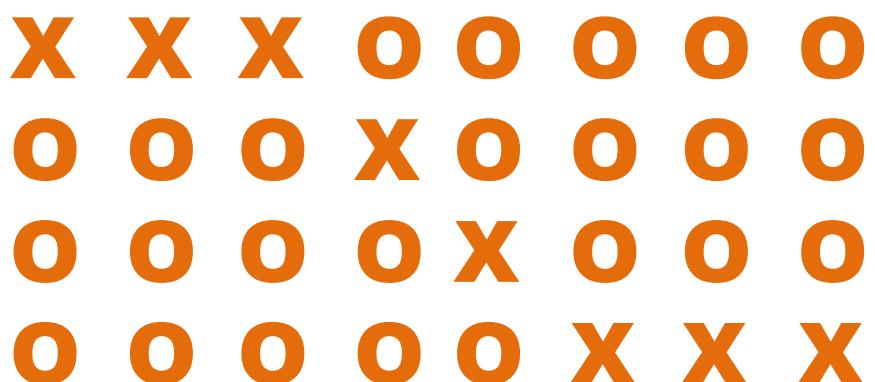
«Gesetz der guten Gestalt»

1. Nähe
2. Ähnlichkeit
3. Geschlossenheit

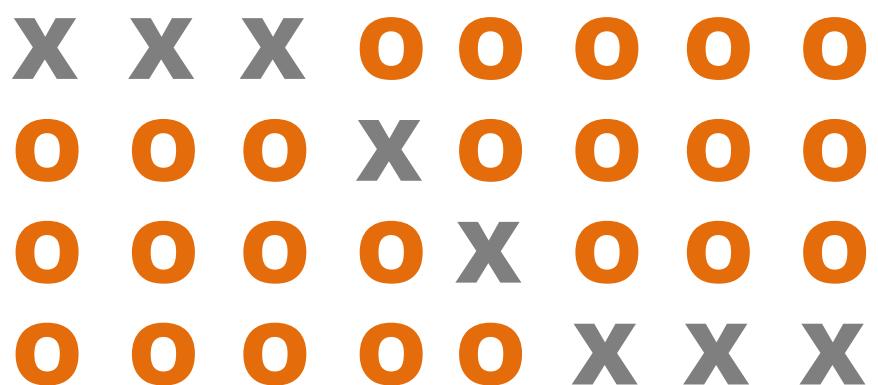


Wieviele Figuren

Kanizsa, G., 1955. Margini quasi-percettivi in campi con stimolazione omogenea, Rivista di Psicologia, 49 (1): 7–30.

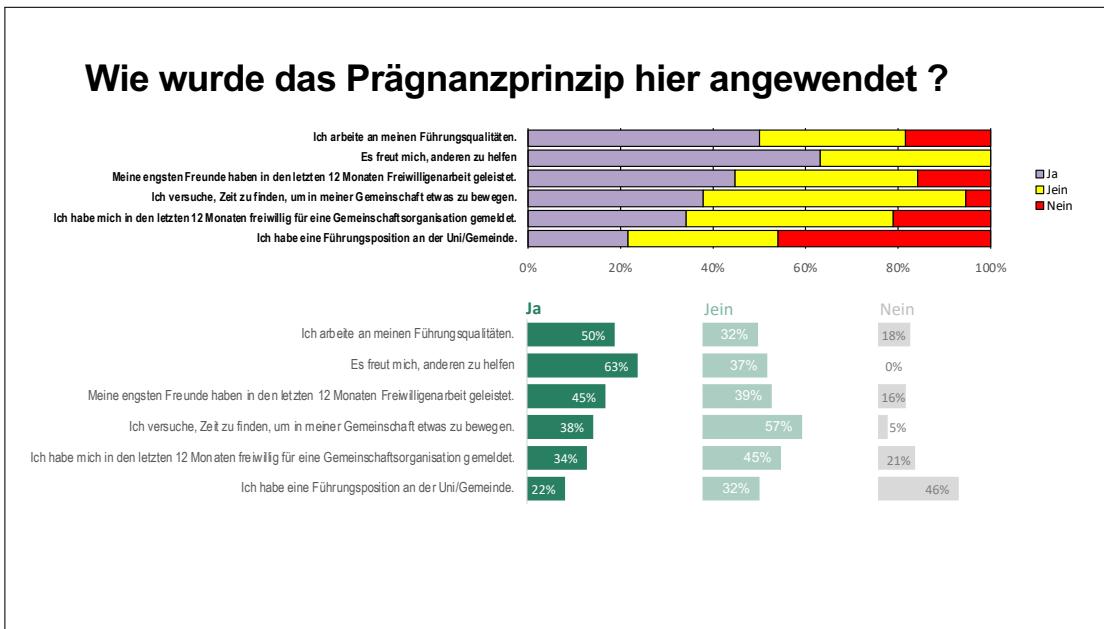
Wieviele Gruppen ?

Wieviele Gruppen ?



«Gesetz der guten Gestalt»

1. Nähe
2. Ähnlichkeit
3. Kontinuität



Referenzen

- Evergreen, Stephanie (2020). Effective Data Visualization. Second Edition, Sage, Thousand Oaks, (CA), U.S.A.
- Gregory, R.L. (1997). Eye and Brain: The Psychology of Seeing (5th ed.). Oxford University Press, Oxford, U.K.
- Ware, C. (2008). Visual Thinking for Design. Morgan Kaufmann, San Francisco, U.S.A.
- Wilke, Claus (2019). Fundamentals of Data Visualization. O'Reilly Media, Inc., Sebastopol, (CA), U.S.A.

Weblinks

<https://serialmentor.com/dataviz/>

<http://persci.mit.edu/gallery/checkershadow>

<https://www.e-study-psychologie.de/mod/book/view.php?id=2929&chapterid=3300>

Referentin

Prof. Sara Irina Fabrikant

Geographisches Institut

© Universität Zürich

Digital Society Initiative



Universität
Zürich UZH

Digital Society Initiative



Studium Digitale

Kursbaustein 7: Effektive Informationsvisualisierung

Lektion 5: **Wozu** visualisiere ich?

Lernziele

- Sie können gezielte Fragen stellen um Mengen, Verteilungen, Anteile, oder Zusammenhänge zu visualisieren.
- Sie sind befähigt den geeigneten Diagrammtyp für Ihre Fragestellung auszuwählen.
- Sie können diverse Visualisierungslösungen vorschlagen, um Mengen, Verteilungen, Anteile oder Zusammenhänge effektiv darzustellen.

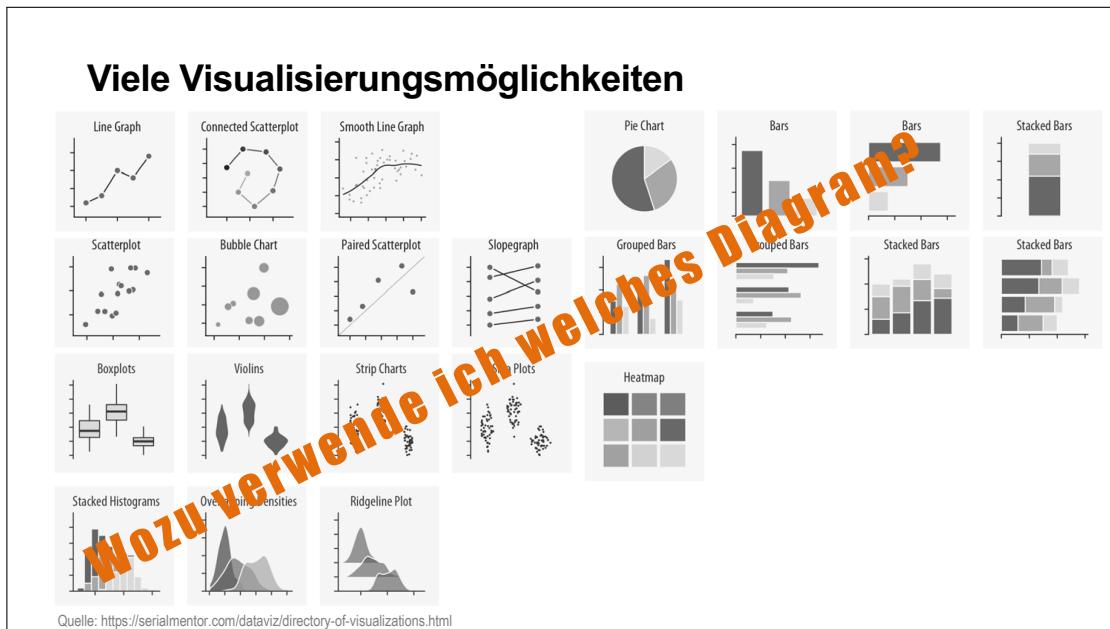
Die 5 Ws in Visualisierung

Warum | Was | für Wen | **Wozu** und Wie wird visualisiert?



Wozu visualisiere ich?

Um die richtige Antwort
auf mein
Informationsbedürfnis zu
erhalten.





Wozu visualisiere ich?

1. Mengen
2. Verteilungen
3. Anteile
4. Zusammenhänge

Visualisierung einer Menge

Wie **hoch** ist der mittlere globale Wasserverbrauch, um 100 Gramm Schokolade herzustellen?



Visualisierung mehrerer Mengen

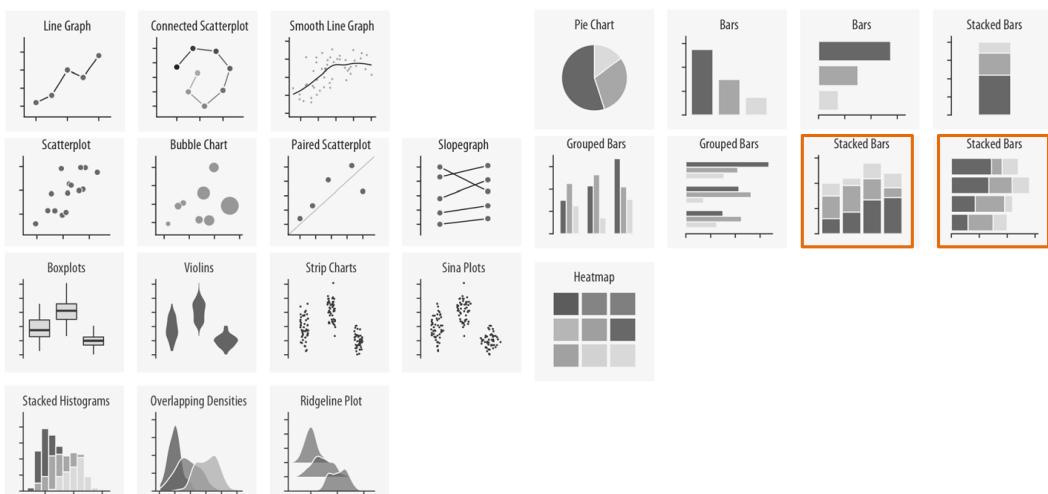
Wie **gross** ist der durchschnittliche virtuelle Wasserfussabdruck von Ländern **im globalen Vergleich**?

Wassermengenvergleich

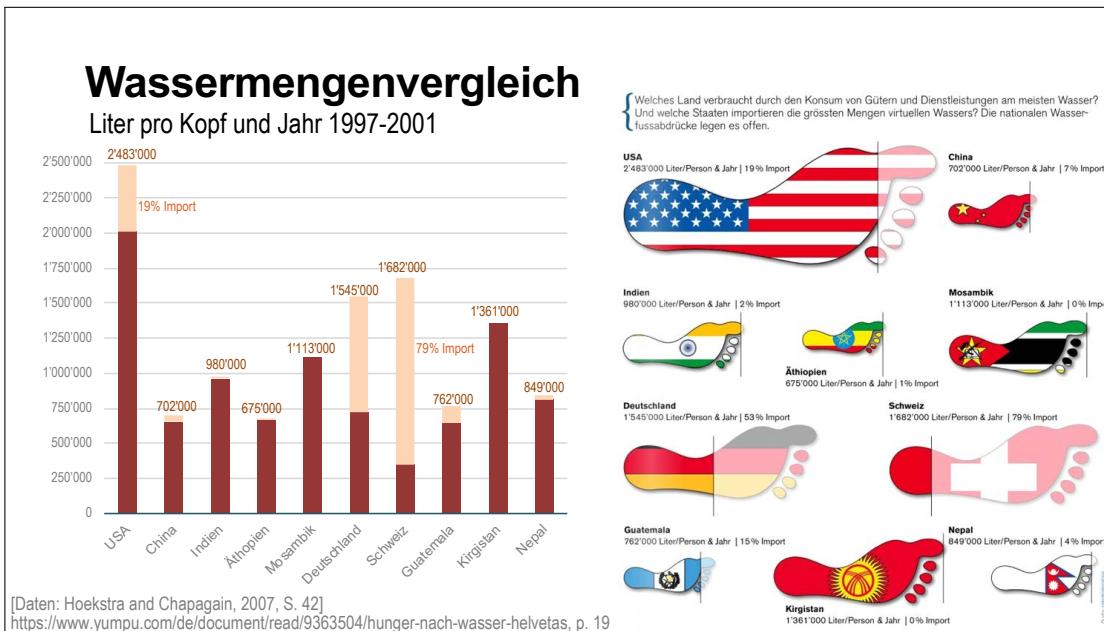
Land	Verbrauch (Liter/Person und Jahr)	Import (%)
USA	2'483'000	19
China	702'000	7
Indien	980'000	2
Äthiopien	675'000	1
Mosambik	1'113'000	0
Deutschland	1'545'000	53
Schweiz	1'682'000	79
Guatemala	762'000	15
Kirgistan	1'361'000	0
Nepal	849'000	4

<https://www.yumpu.com/de/document/read/9363504/hunger-nach-wasser-helvetas>

Viele Visualisierungsmöglichkeiten



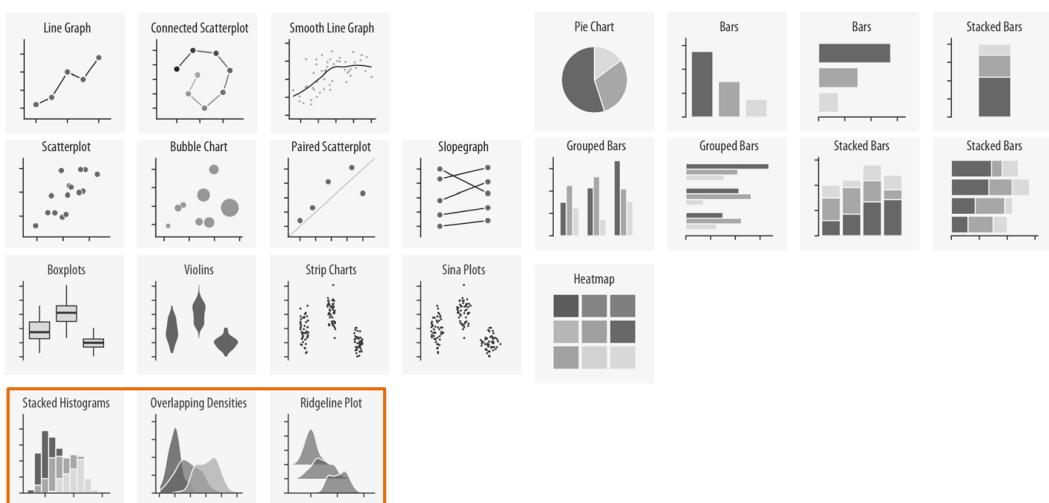
Quelle: <https://serialmentor.com/dataviz/directory-of-visualizations.html>

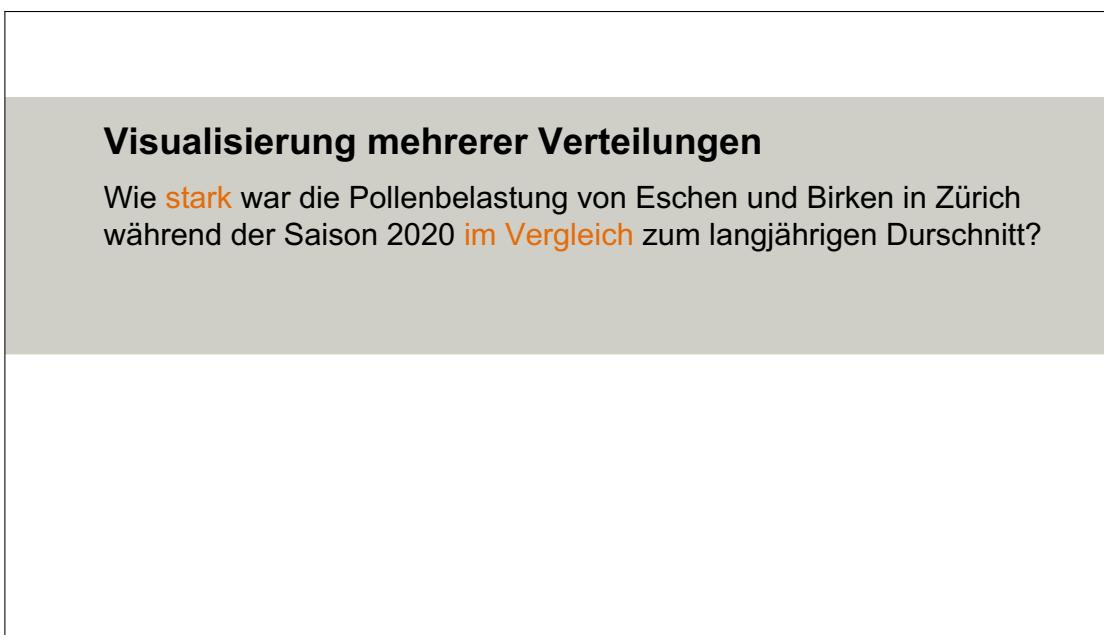
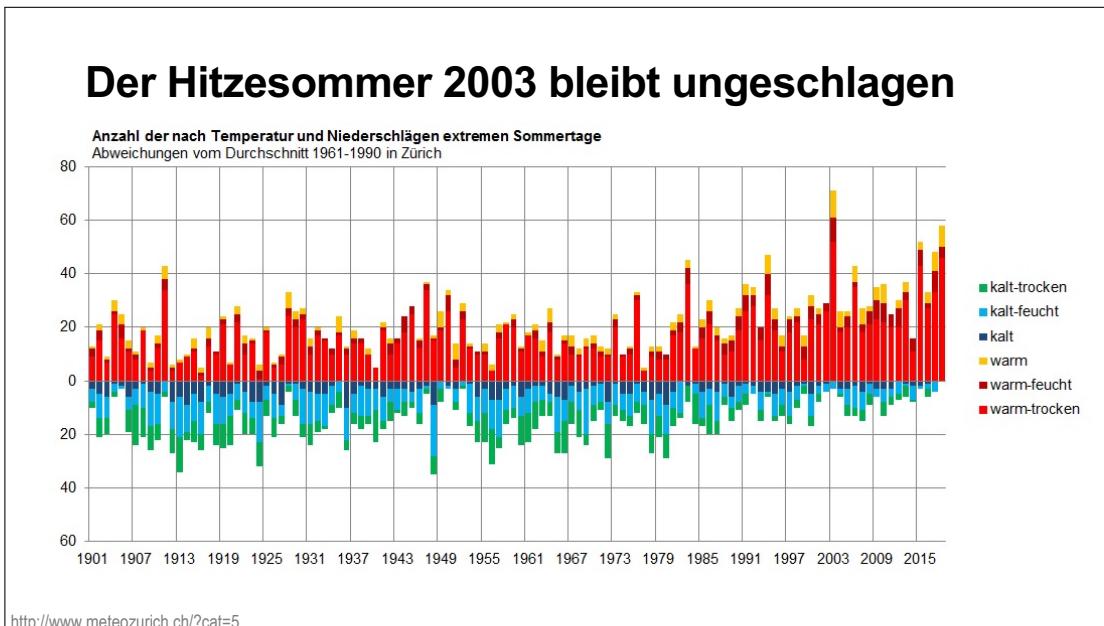


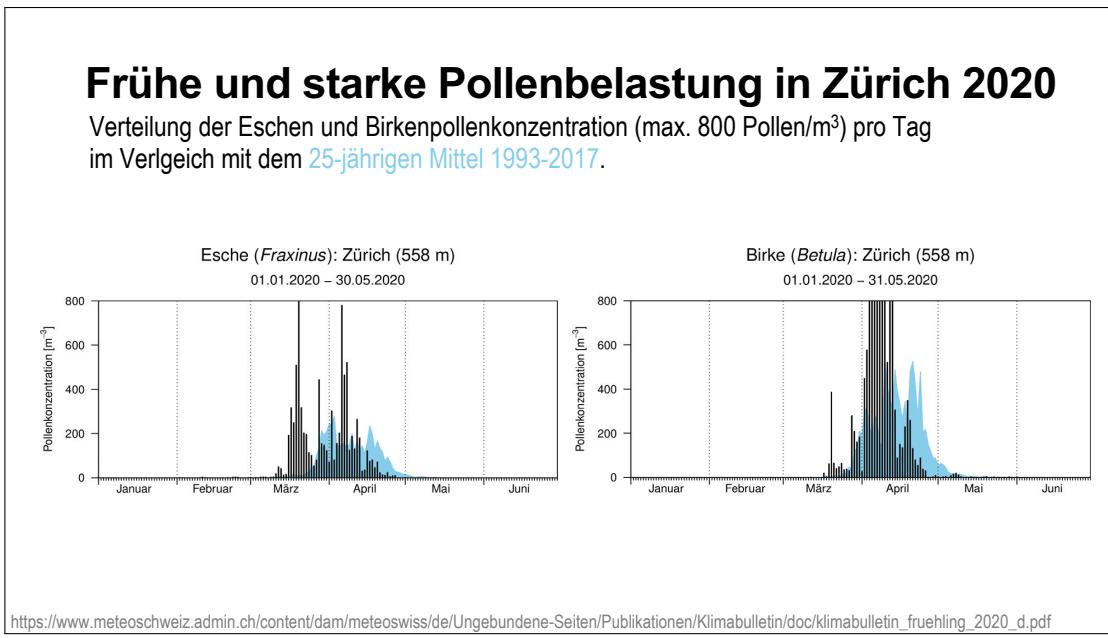
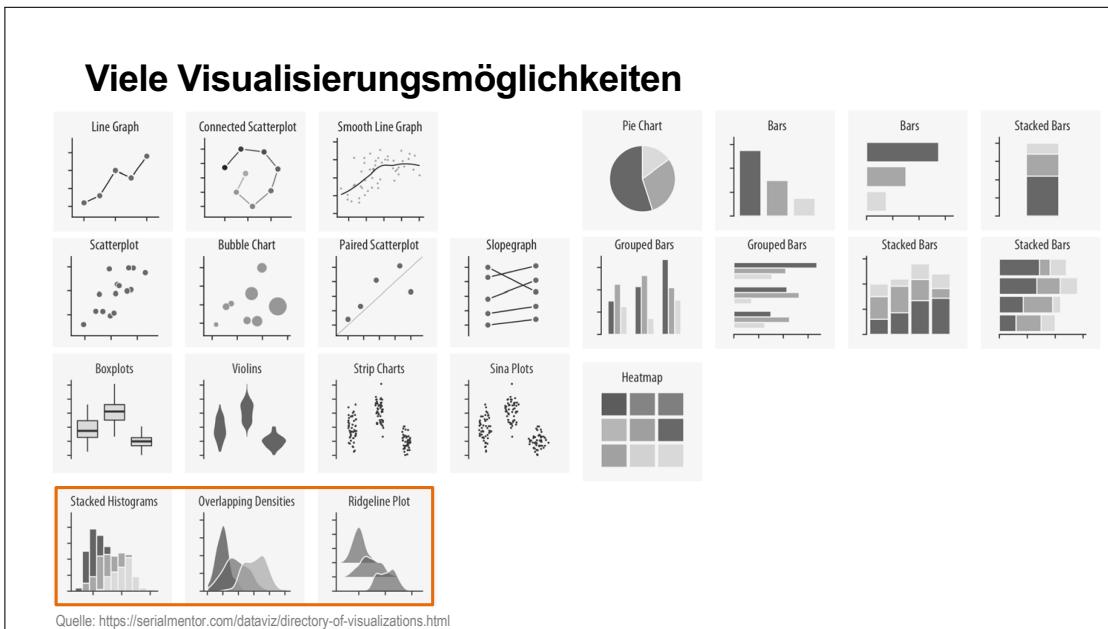
Visualisierung einer Verteilung

Wird der Sommer immer **heisser** in der Stadt Zürich?

Viele Visualisierungsmöglichkeiten







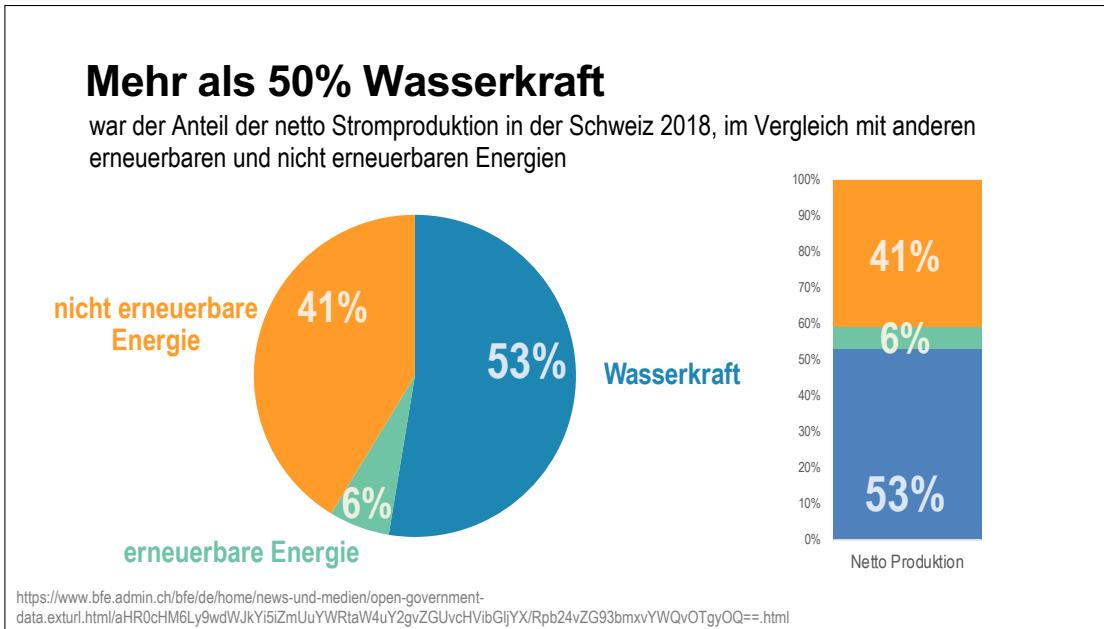
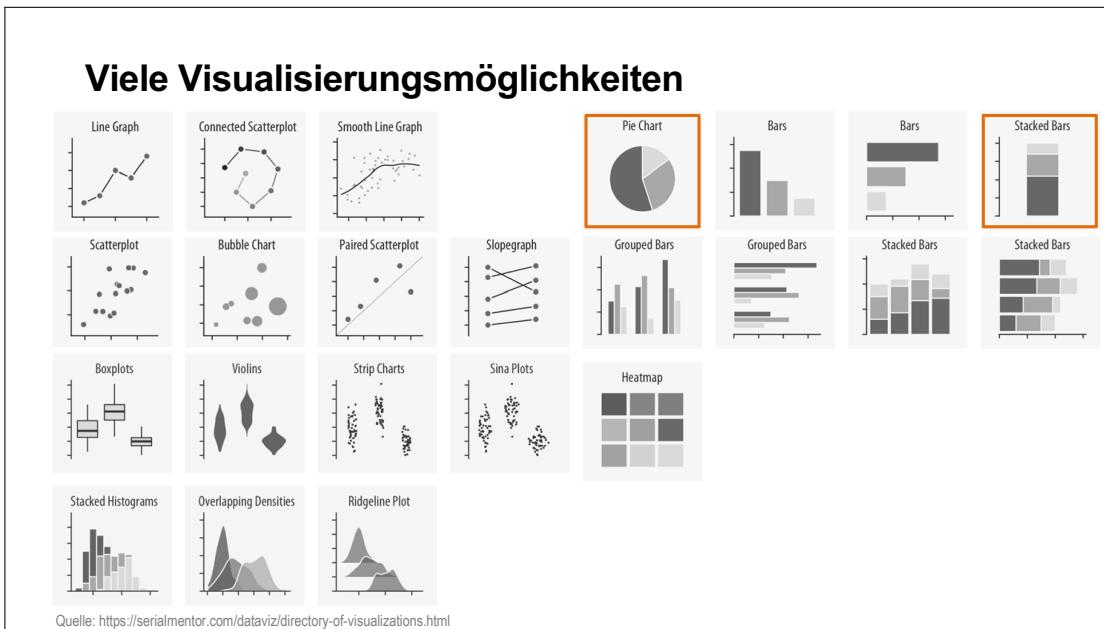


Wozu visualisiere ich?

- 1. Mengen
- 2. Verteilungen
- 3. Anteile**
- 4. Zusammenhänge

Visualisierung eines Anteils

Wie **gross** ist der **Anteil** der Wasserkraft an der **gesamten** Stromproduktion in der Schweiz?

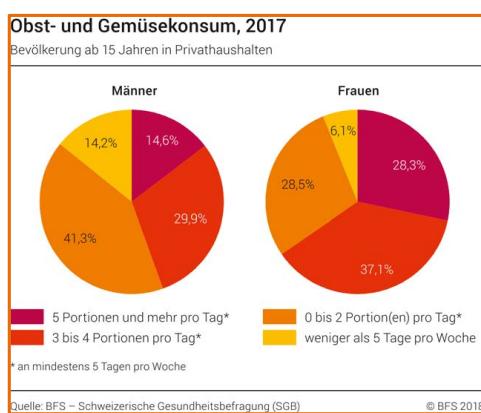


Visualisierung mehrerer Anteile

Wie gross ist der **Anteil** der Personen in der Schweiz, die mindestens fünf Portionen Obst oder Gemüse pro Woche konsumieren? Gibt es Geschlechterunterschiede?

Doppelt soviele Frauen wie Männer

konsumierten in der Schweiz mindestens 5 Portionen Obst- und Gemüse pro Woche in 2017.



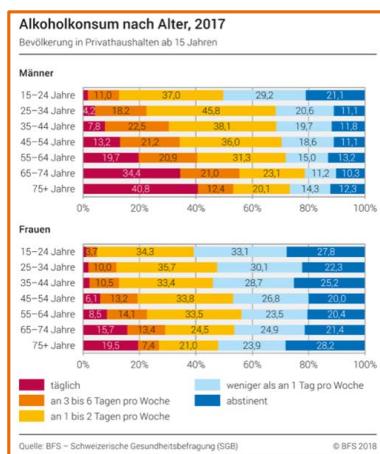
<https://www.bfs.admin.ch/bfs/de/home/statistiken/gesundheit/determinanten/ernaehrung.html>

Visualisierung mehrerer Anteile

Welche Altersklasse hat den **höchsten Anteil** am täglichen Alkoholkonsum in der Schweiz. Gibt es Geschlechterunterschiede?

Doppelt soviele Männer ab 75 J. wie Frauen

trinken täglich Alkohol in der Schweiz 2017



<https://www.bfs.admin.ch/bfs/de/home/statistiken/gesundheit/determinanten/alkohol.html>

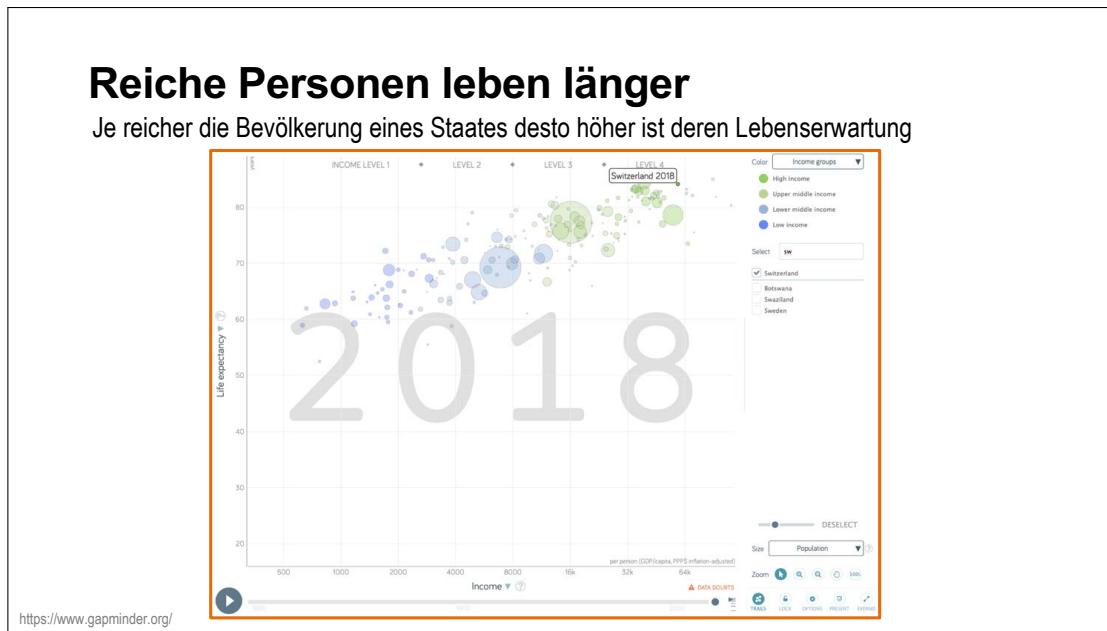
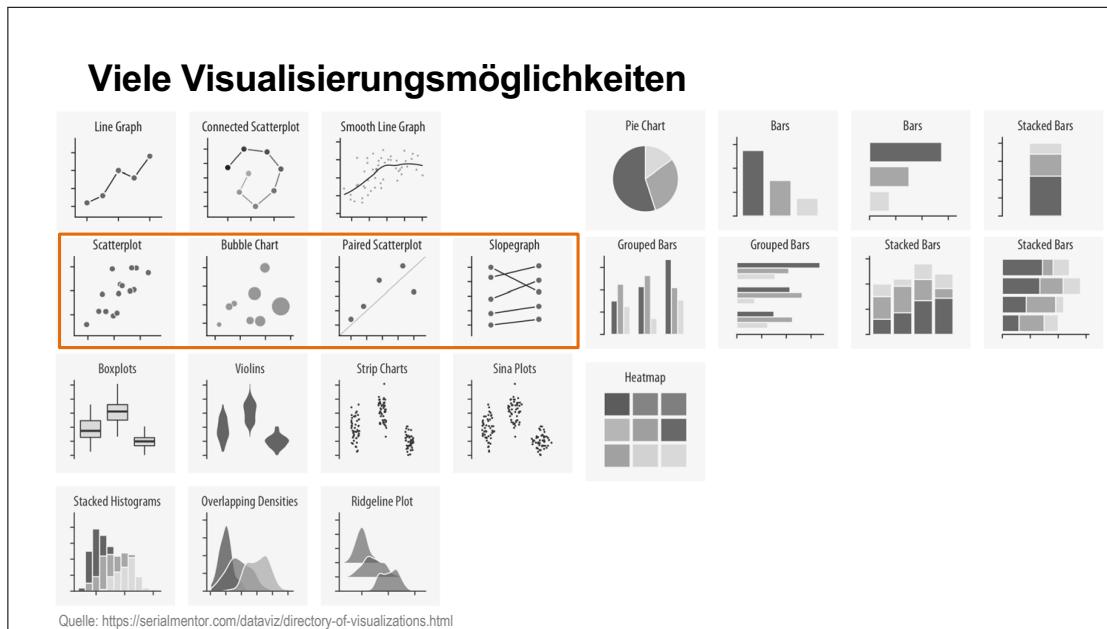


Wozu visualisiere ich?

1. Mengen
2. Verteilungen
3. Anteile
4. **Zusammenhänge**

Visualisierung eines Zusammenhangs

Besteht ein **Zusammenhang** zwischen dem Einkommen und der Lebenserwartung in der Weltbevölkerung?

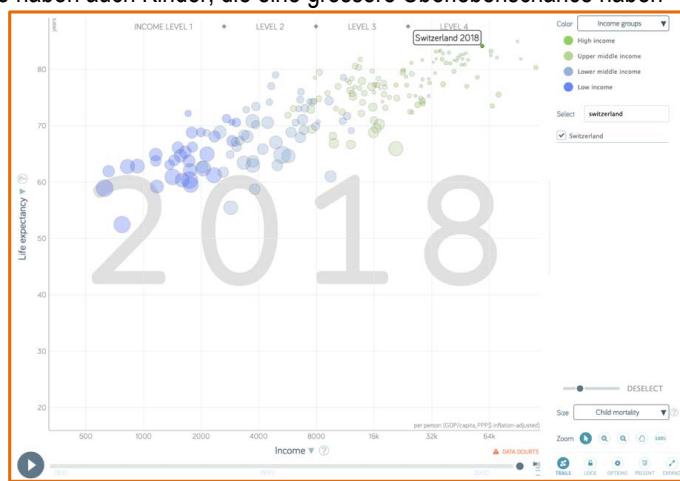


Visualisierung von mehreren Zusammenhängen

Besteht ein **Zusammenhang** zwischen dem Einkommen, der Lebenserwartung und der Kindersterblichkeit der Weltbevölkerung?

Reiche Personen leben nicht nur länger

sondern sie haben auch Kinder, die eine grössere Überlebenschance haben



<https://www.gapminder.org/>

Referenzen

Evergreen, Stephanie (2020). Effective Data Visualization. Second Edition, Sage, Thousand Oaks, (CA), U.S.A.

Hoekstra, A.Y., Chapagain, A.K. Water footprints of nations: Water use by people as a function of their consumption pattern. *Water Resour Manage* 21, 35–48 (2007). <https://doi.org/10.1007/s11269-006-9039-x>, Table 3, p. 42

Wilke, Claus (2019). Fundamentals of Data Visualization. O'Reilly Media, Inc., Sebastopol, (CA), U.S.A.

Weblinks

<https://serialmentor.com/dataviz/>

https://www.pseau.org/outils/ouvrages/waterfootprint_comprehensive_introduction_to_water_footprints_en.pdf

http://www.education21.ch/sites/default/files/uploads/pdf-d/Mystery_Virtuelles-Wasser_Sek-II_de.pdf

Rosling TED Talk 2006:

https://www.ted.com/talks/hans_rosling_the_best_stats_you_ve_ever_seen#t-267880

Referentin

Prof. Sara Irina Fabrikant
Geographisches Institut

© Universität Zürich

Digital Society Initiative



Universität
Zürich UZH

Digital Society Initiative



Studium Digitale

Kursbaustein 7: Effektive Informationsvisualisierung

Lektion 5: Wie visualisiere ich?

Lernziele

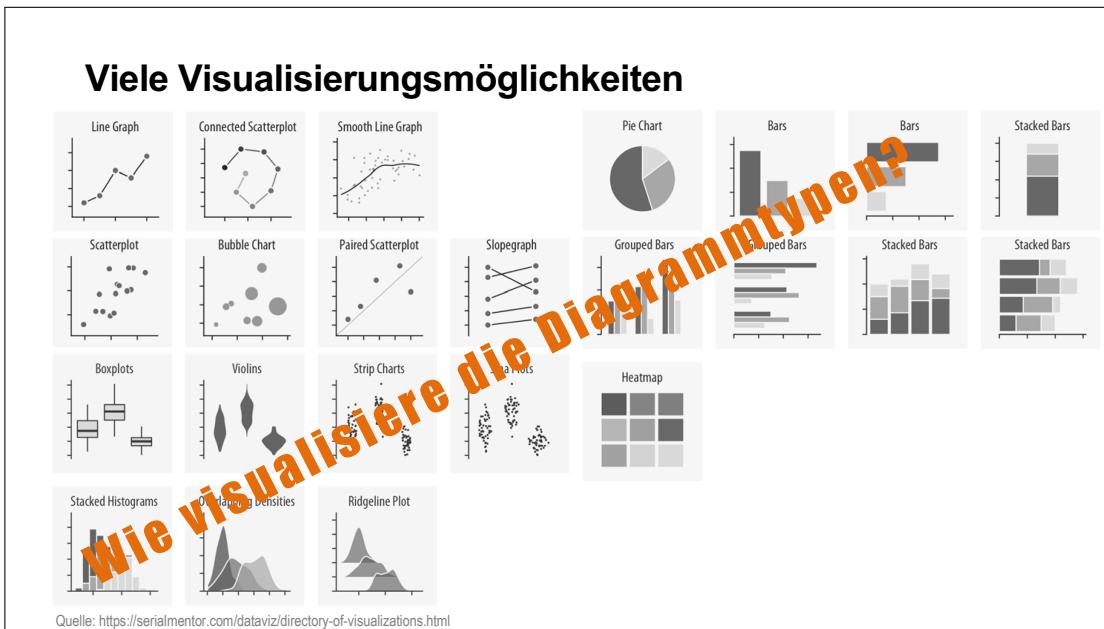
- Sie wissen welche Diagrammkomponenten bearbeitet werden sollten, um die Lesbarkeit der Visualisierung zu erhöhen.
- Sie kennen die graphischen Variablen, um Diagrammtypen attraktiv und gut lesbar zu gestalten
- Sie können diverse graphische Lösungen vorschlagen, um Mengen, Verteilungen, Anteile oder Zusammenhänge prägnant darzustellen.

Die 5 Ws in Visualisierung

Warum | Was | für Wen | Wozu und Wie wird visualisiert?

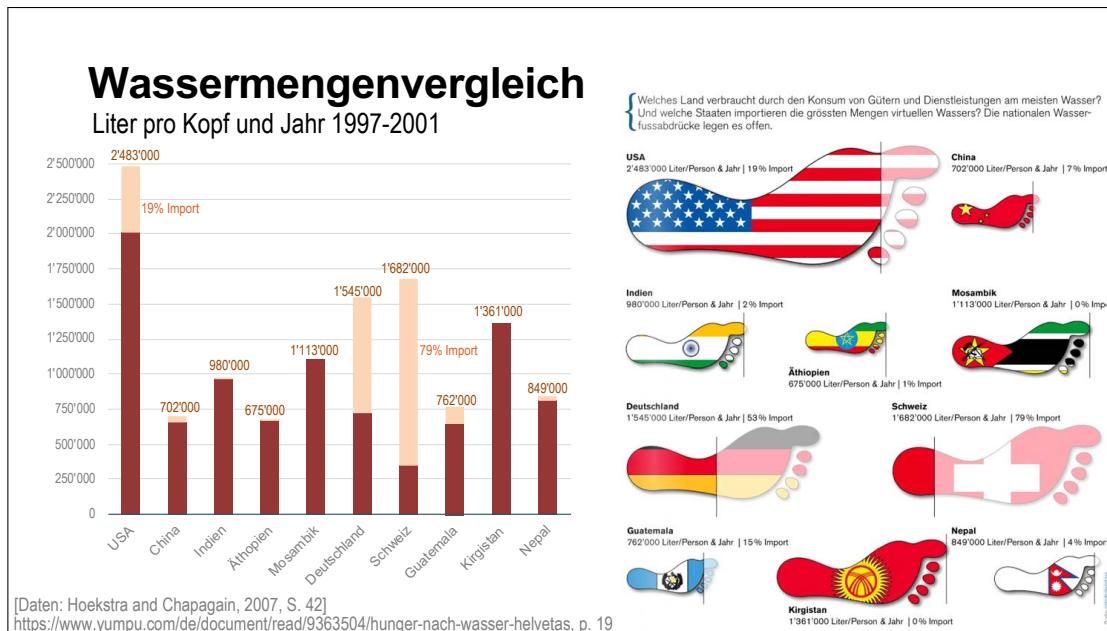
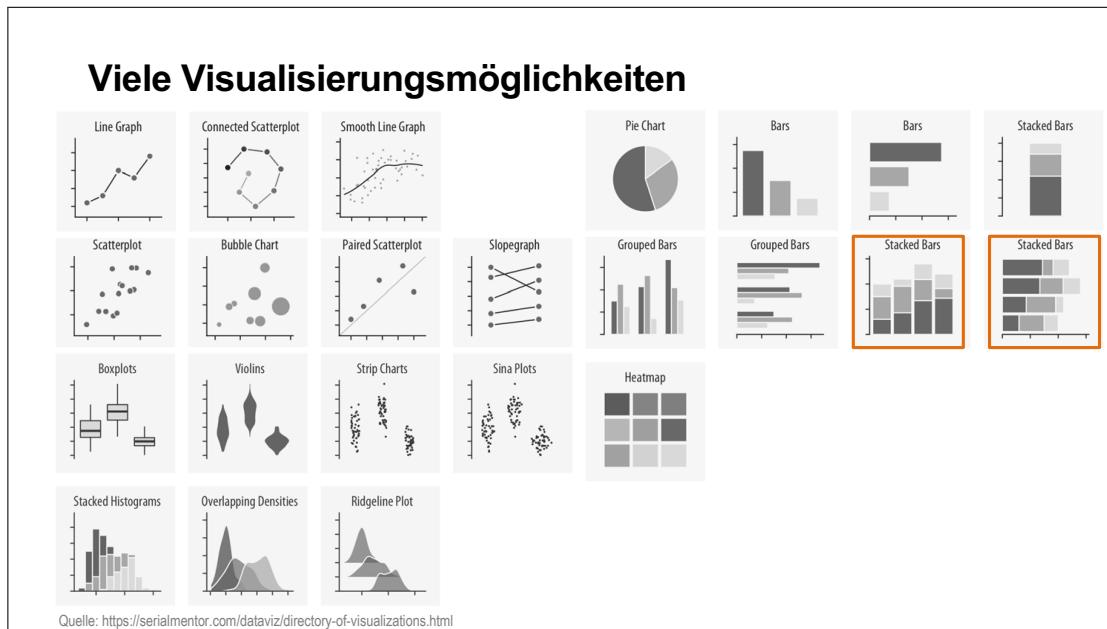
Wie visualisiere ich?





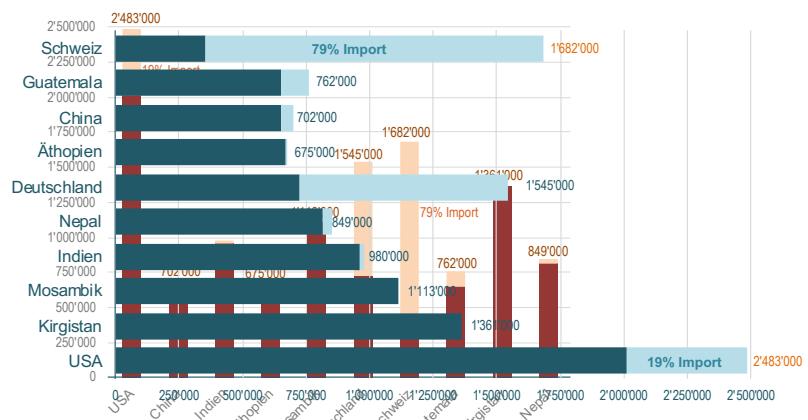
Wie visualisiere ich?

1. Mengen
2. Verteilungen
3. Anteile
4. Zusammenhänge



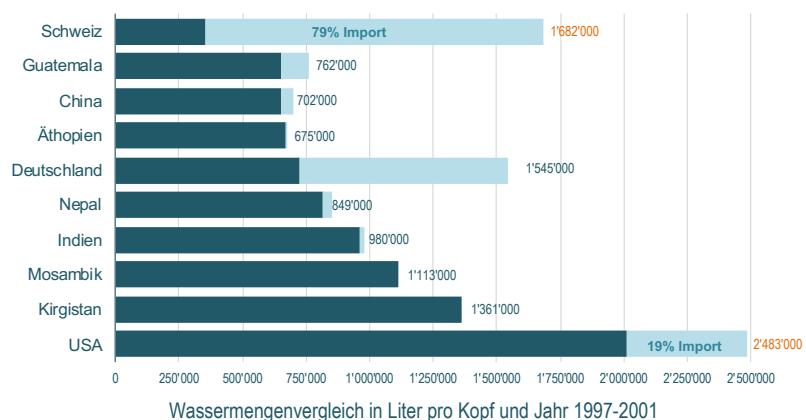
Wassermengenvergleich

Liter pro Kopf und Jahr 1997-2001



[Daten: Hoekstra and Chapagain, 2007, S. 42]
<https://www.yumpu.com/de/document/read/9363504/hunger-nach-wasser-helvetas, p. 19>

Die Schweiz importiert 79% ihres Wasserfußabdrucks



[Daten: Hoekstra and Chapagain, 2007, S. 42]
<https://www.yumpu.com/de/document/read/9363504/hunger-nach-wasser-helvetas, p. 19>

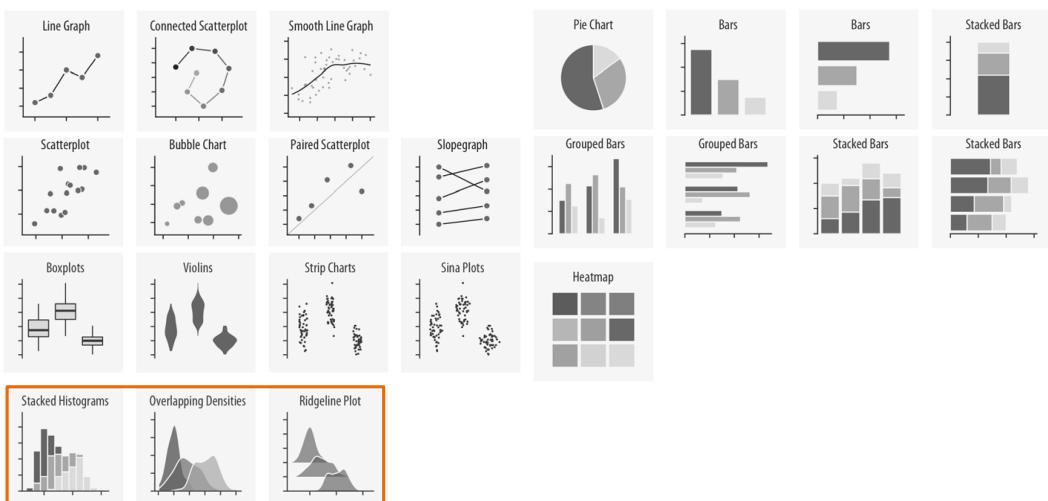
Nun kommen wir zu graphischen Verbesserungsvorschlägen, um Verteilungen prägnant zu visualisieren



Wozu visualisiere ich?

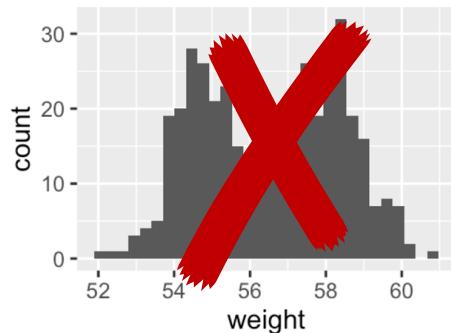
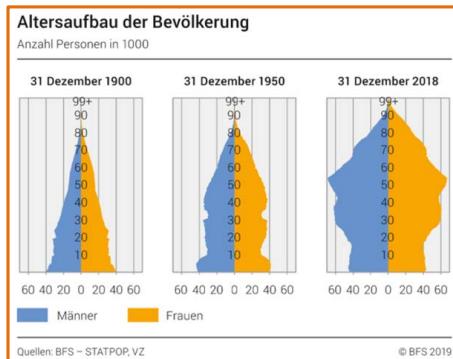
1. Mengen
2. **Verteilungen**
3. Anteile
4. Zusammenhänge

Viele Visualisierungsmöglichkeiten

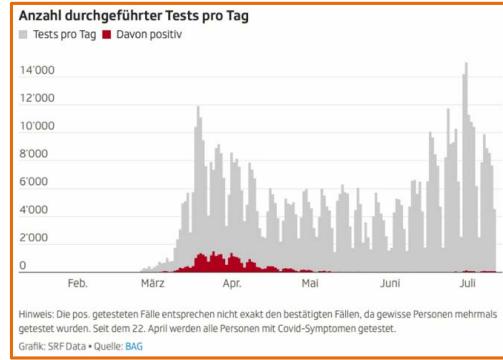
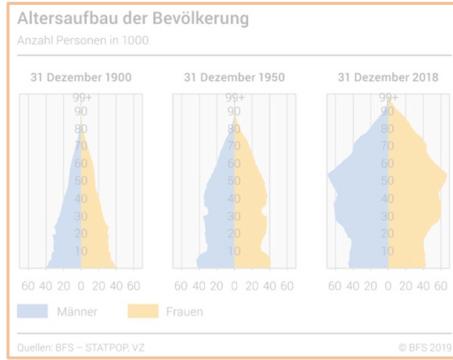


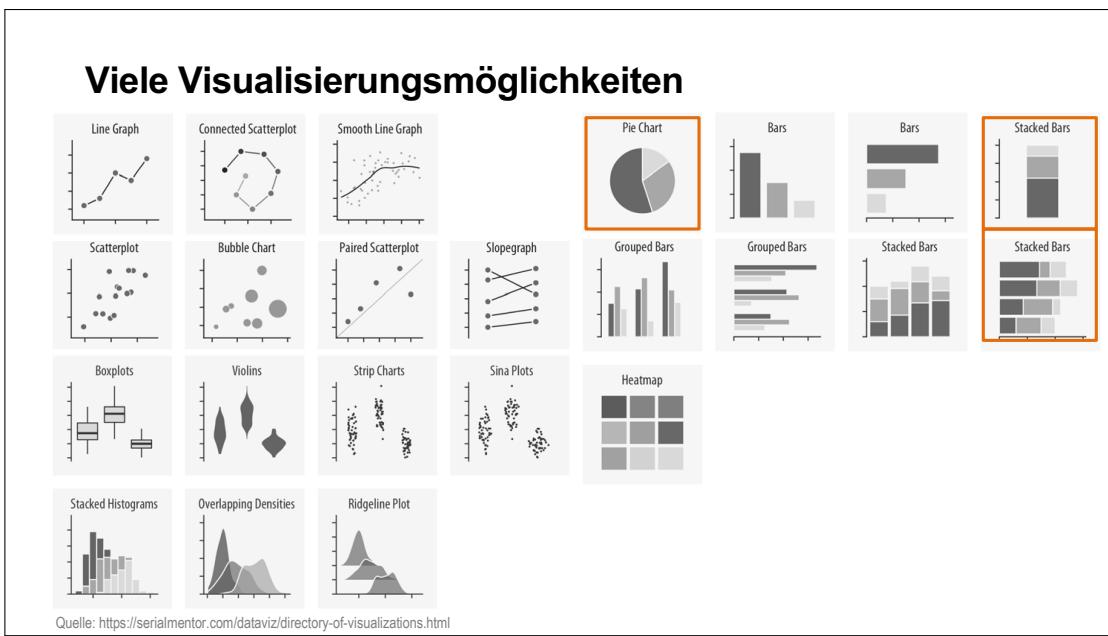
Quelle: <https://serialmentor.com/dataviz/directory-of-visualizations.html>

Die Schweizer Bevölkerung wird immer älter



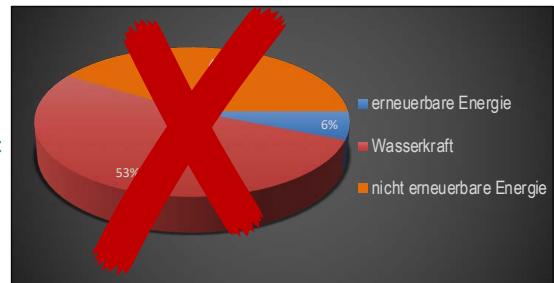
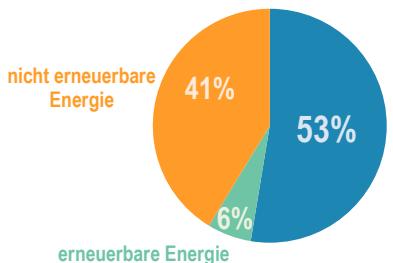
Die positiv getesteten Corona Fälle sind stark gesunken, obwohl insgesamt mehr getestet wird





Mehr als 50% Wasserkraft

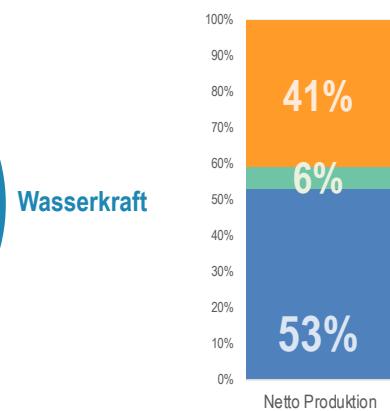
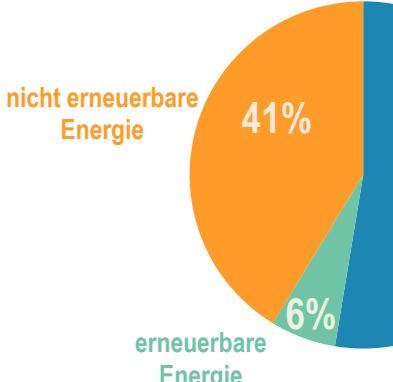
war der Anteil der netto Stromproduktion in der Schweiz 2018, im Vergleich mit anderen erneuerbaren und nicht erneuerbaren Energien



<https://www.bfe.admin.ch/bfe/de/home/news-und-medien/open-government-data/exturi.html/aHR0cHM6Ly9wdWJkYi5ZmUuYWRTaW4uY2gvZGUvcHViGjYX/Rpb24vZG93bmvxYWQvOTgyOQ==.html>

Mehr als 50% Wasserkraft

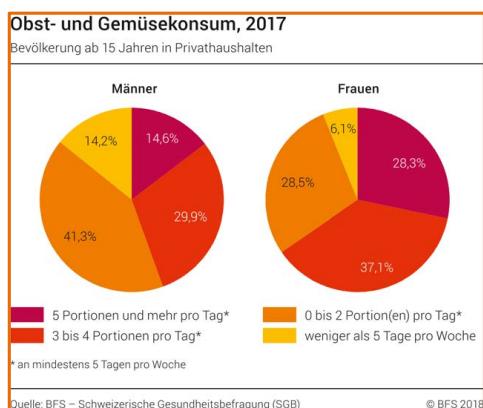
war der Anteil der netto Stromproduktion in der Schweiz 2018, im Vergleich mit anderen erneuerbaren und nicht erneuerbaren Energien



<https://www.bfe.admin.ch/bfe/de/home/news-und-medien/open-government-data/exturi.html/aHR0cHM6Ly9wdWJkYi5ZmUuYWRTaW4uY2gvZGUvcHViGjYX/Rpb24vZG93bmvxYWQvOTgyOQ==.html>

Doppelt soviele Frauen wie Männer

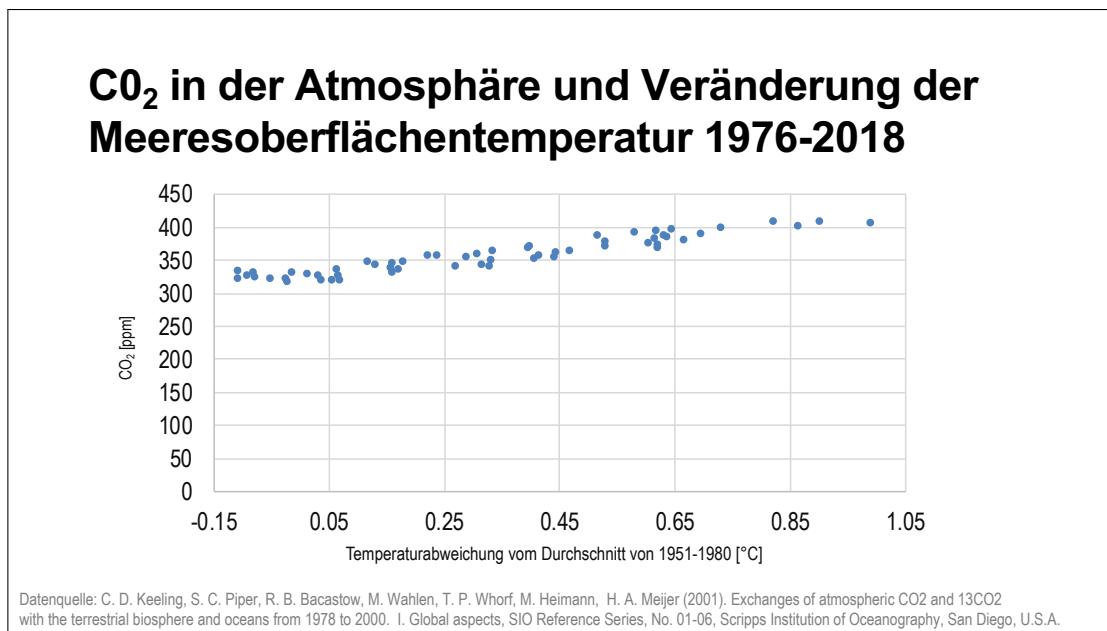
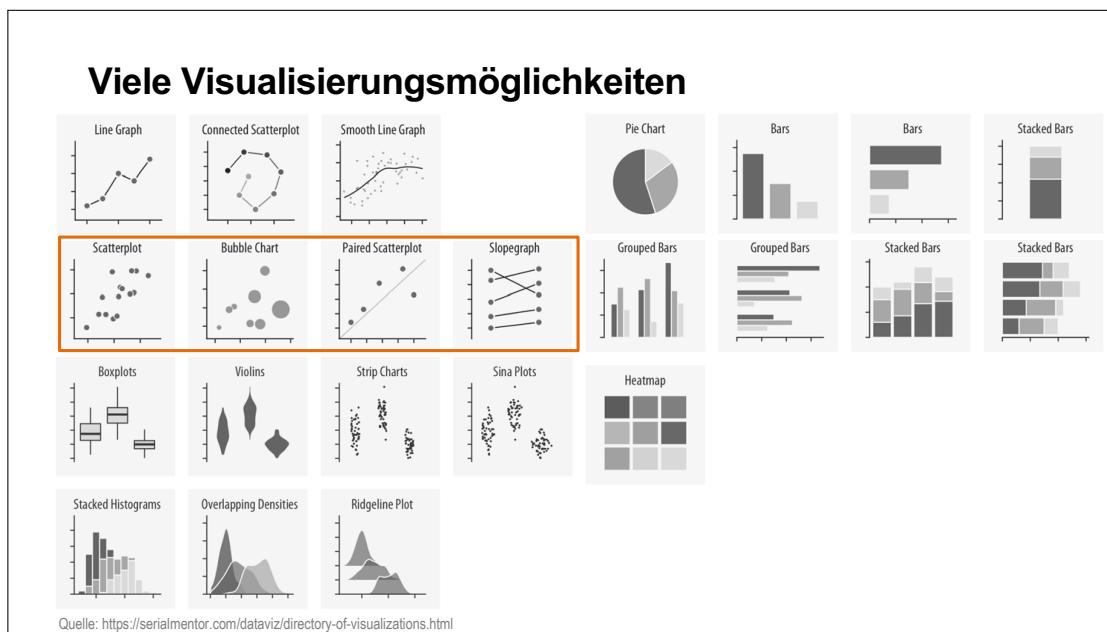
konsumierten in der Schweiz mindestens 5 Portionen Obst- und Gemüse pro Woche in 2017.



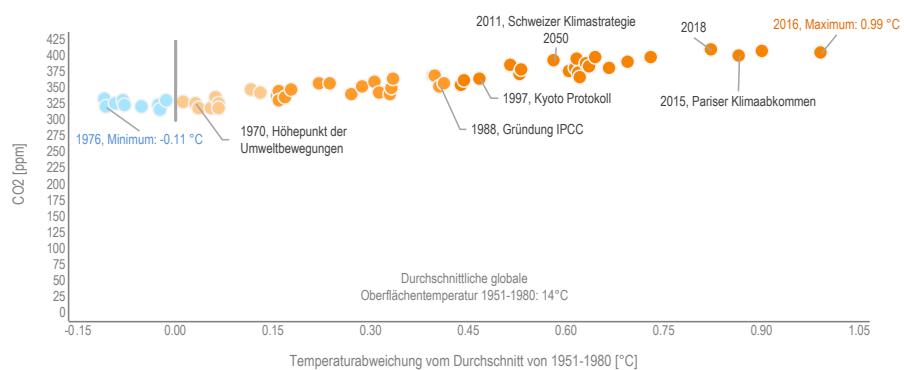
<https://www.bfs.admin.ch/bfs/de/home/statistiken/gesundheit/determinanten/ernaehrung.html>

Wozu visualisiere ich?

1. Mengen
2. Verteilungen
3. Anteile
4. Zusammenhänge

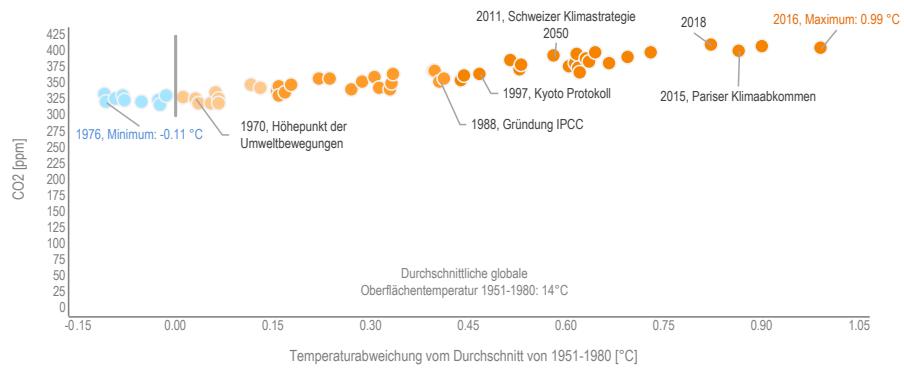


Je mehr CO₂ in der Luft, desto wärmer das Meer



Datenquelle: C. D. Keeling, S. C. Piper, R. B. Bacastow, M. Wahlen, T. P. Whorf, M. Heimann, H. A. Meijer (2001). Exchanges of atmospheric CO₂ and ¹³CO₂ with the terrestrial biosphere and oceans from 1978 to 2000. I. Global aspects, SIO Reference Series, No. 01-06, Scripps Institution of Oceanography, San Diego, U.S.A.

Je mehr CO₂ in der Luft, desto wärmer das Meer



Datenquelle: C. D. Keeling, S. C. Piper, R. B. Bacastow, M. Wahlen, T. P. Whorf, M. Heimann, H. A. Meijer (2001). Exchanges of atmospheric CO₂ and ¹³CO₂ with the terrestrial biosphere and oceans from 1978 to 2000. I. Global aspects, SIO Reference Series, No. 01-06, Scripps Institution of Oceanography, San Diego, U.S.A.

Referenzen

Evergreen, Stephanie (2020). Effective Data Visualization. Second Edition, Sage, Thousand Oaks, (CA), U.S.A.

Hoekstra, A.Y., Chapagain, A.K. Water footprints of nations: Water use by people as a function of their consumption pattern. Water Resour Manage 21, 35–48 (2007).
<https://doi.org/10.1007/s11269-006-9039-x>, Table 3, p. 42

Wilke, Claus (2019). Fundamentals of Data Visualization. O'Reilly Media, Inc., Sebastopol, (CA), U.S.A.

Weblinks

<https://serialmentor.com/dataviz/>

<https://stephanieevergreen.com/books/>

https://www.pseau.org/outils/ouvrages/waterfootprint_comprehensive_introduction_to_water_footprints_en.pdf

http://www.education21.ch/sites/default/files/uploads/pdf-d/Mystery_Virtuelles-Wasser_Sek-II_de.pdf

<https://vega.github.io>

Referentin

Prof. Sara Irina Fabrikant
Geographisches Institut

© Universität Zürich
Digital Society Initiative

5 Datenmanagement

SVEN HELMER

Studium Digitale Kursbaustein 3: Datenmanagement

Begleit-Skript Dieses Skript bereitet die Inhalte des Kursbaustein 3 schriftlich auf. Die Inhalte sind nicht identisch mit den Videos; sie wurden angepasst für die Lesbarkeit als Dokument.

5.1 Lektion 1: Wieso eigentlich Datenmanagement?

Die Menschheit produziert tagtäglich Unmengen von Daten, sei es auf Social-Network-Plattformen, in Logdateien auf Webservern (die unser Surfverhalten mitprotokollieren), in Sensoren die inzwischen an vielen Maschinen und Fahrzeugen angebracht sind oder in der Protokollierung von Lieferketten. Schätzungen gehen davon aus, dass wir im Jahr 2025 jeden Tag ca. 463 Exabytes an Daten erzeugen werden, das entspricht dem Inhalt von ungefähr 212 Milliarden DVDs.

Eine derart grosse Menge an Daten speichert man nicht mehr einfach in Dateien auf einem Rechner, das Ganze muss professioneller organisiert werden. Um dieses Thema geht es in diesem Kursbaustein: wie verwalte ich Daten, so dass ich sie später effektiv einsetzen kann?

Datenmanagement kann kurz so definiert werden:

- Es geht um das Sammeln, Speichern und Verarbeiten von Daten
- und das Ganze sollte zuverlässig, sicher und effizient geschehen.

Leider wird mit Daten oft nicht sorgfältig umgegangen, was dazu führt, dass sie später entweder unbrauchbar werden oder mit viel Aufwand rekonstruiert und aufbereitet werden müssen. Eventuell gehen sie sogar ganz verloren.

5.2 Lektion 2: Lebenszyklus von Daten

Die meisten Daten durchlaufen einen typischen Lebenszyklus und leben oft länger als man denkt. Oft werden die Daten auch von anderen Benutzern übernommen und weiterverwendet. Wenn unsere Daten schlecht organisiert und beschrieben sind, wird das schwierig. Mehr über die korrekte Beschreibung von

Daten und Konventionen zur Namensgebung erfahren Sie in einem späteren Kursbaustein.

Wie sieht ein typischer Lebenszyklus von Daten (englisch: data lifecycle) eigentlich aus? Wir schauen uns die wichtigsten Phasen einmal an:

- Phase 1 (Erzeugen): zunächst werden die Daten gesammelt bzw. erzeugt. Das können Logdateien auf einem Webserver sein der Zugriffe mitprotokolliert oder experimentelle Daten die von Sensoren ausgegeben werden. Hier macht es Sinn sich Gedanken zu machen, welche Daten konkret gebraucht werden und wie die Daten organisiert und strukturiert sind (wir kommen auf die Strukturierung von Daten in der Lektion 5 über Datenmodellierung zurück).
- Phase 2 (Vorverarbeiten): oft gibt es Probleme mit der Qualität der gesammelten Daten, z.B. durch falsche Eingaben oder gestörte Sensoren. Die Daten müssen erst einmal gesichtet und eventuell überarbeitet werden, um ihre Qualität zu verbessern. Sie lernen Verfahren aus der Datenanalyse und der Statistik die beim Aufspüren von Fehlern helfen können in einem späteren Kursbaustein kennen.
- Phase 3 (Speichern): wie speichern wir die Daten möglichst geschickt ab, damit sie optimal genutzt werden können? Das bedeutet konkret, dass ich genau das finde was ich suche und das möglichst schnell. In Lektion 4 schauen wir uns an, wie Datenbanksysteme zur Speicherung eingesetzt werden können.
- Phase 4 (Nutzung): Jetzt kann die eigentliche Nutzung der Daten erfolgen. Da dies stark von der Anwendung abhängt, schauen wir uns hier nur ein paar Beispiele an. Lektion 6 zeigt, wie wir auf Daten die in einem Datenbanksystem abgespeichert sind, zugreifen können.
- Phase 5 (Archivierung/Löschen): wenn die Daten nicht mehr aktiv verwendet werden, steht die Archivierung an. Falls benötigt können sie dann zu einem späteren Zeitpunkt wieder abgerufen werden. Es kann aber auch sein, dass die Daten gelöscht werden müssen. Aus rechtlichen Gründen dürfen bestimmte Daten nur eine gewisse Zeit aufbewahrt werden. Mehr zu diesem Thema gibt es in einem anderen Kursbaustein.

5.3 Lektion 3: Spreadsheets

Spreadsheets, neudeutsch für Tabellenkalkulation, war eine der ersten Killer-Apps die für PCs entwickelt wurden. Im Prinzip ist es eine elektronische Umset-

zung von Vorgängen die man aus der Buchhaltung kennt und es ist ein Werkzeug für die Eingabe, Speicherung, Analyse und Visualisierung von Daten. Dabei besteht ein Spreadsheet aus einer Tabelle mit Zeilen und Spalten:

- Die oberste Zeile enthält Beschreibungen der Spalten
- Jede Zeile darunter enthält einen Datensatz
- Jede Spalte steht dabei für eine Variable
- Jede Zelle enthält einen Wert. Dieser Wert kann auch aus anderen Zellwerten berechnet werden.

Die folgende Abbildung zeigt ein Beispiel eines Spreadsheets mit Informationen über die Passagierfrequenz verschiedener Bahnhöfe der SBB.

	A	B	C	D	E	F	G	H
1	Code	Bahnhof Haltestelle	Bezugsjahr	Kanton	DTV	DWV	DNWV	
2	AD	Aadorf	2018	TG	1700	2000	1000	
3	AA	Aarau	2018	AG	37900	44800	22700	
4	ABO	Aarburg-Ottringen	2018	AG	2500	3000	1300	
5	AAT	Aathal	2018	ZH	740	800	610	
6	ACLA	Acla da Fontauna	2018	GR	90	90	80	
7	AE	Aesch	2018	BL	2000	2400	1300	
8	AF	Affoltern am Albis	2018	ZH	6800	8100	4000	
9	AIG	Aigle	2018	VD	6800	7700	5000	
10	AI	Airolo	2018	TI	420	370	510	

Die erste Zeile im Spreadsheet sagt uns, dass die Spalten Informationen über den Bahnhofscode, den Namen des Bahnhofs, das Jahr der Datenerhebung, den Kanton in dem sich der Bahnhof befindet und die Passagierfrequenzen enthält. Dabei steht DTV für durchschnittlicher täglicher Verkehr, DWV für durchschnittlicher Wochentagsverkehr und DNWV für durchschnittlicher Nichtwochentagsverkehr. Die Information über den ersten Bahnhof in der zweiten Zeile des Spreadsheets kann man in natürlicher Sprache folgendermassen zusammenfassen: der Bahnhof Aadorf mit dem Code AD im Kanton Thurgau hatte für das Jahr 2018 folgendes Passagieraufkommen: DTV – 1700, DWV – 2000, DNWV – 1000.

Die Best Practices für den Umgang mit Daten die in einem späteren Kursbaustein detaillierter vorgestellt werden (wie z.B. sinnvolle Namensgebung von Dateien, die Speicherung von Metadaten mit den eigentlichen Daten, das Anlegen von Sicherungskopien, sowie die Verwendung von nicht-proprietären Dateiformaten), gelten auch für Spreadsheets. Zusätzlich dazu sollte man bei Spreadsheets noch folgendes beachten:

- Nur einen Wert pro Zelle eintragen.
- Zellen nicht einfach leer lassen. Sonst weiss man später nicht ob der Wert fehlt, weil es ihn nicht gibt oder ob er nicht bekannt ist.
- Man sollte keine Berechnungen auf den Rohdaten ausführen die die Rohdaten ändern. Solche Berechnungen immer auf Kopien ausführen.
- Um unsinnige Dateneingabe zu verhindern, sollte man die Datenvalidierungsfunktionen von Spreadsheets nutzen: z.B. Datentypen festlegen oder Wertebereiche überprüfen.
- Schauen, ob die Modellierung sinnvoll ist indem man eine Zeile wie oben für den Bahnhof Aadorf in natürlicher Sprache ausdrückt.

Spreadsheets haben eine Reihe von Vorteilen und Nachteilen, schauen wir uns erst die Vorteile an:

- Es ist relativ einfach Daten einzutragen und abzuspeichern.
- Daten können in gängigen Formaten importiert und exportiert werden.

Es gibt aber auch eine Reihe von Nachteilen (wie die ausgebessert werden, sehen wir in der nächsten Lektion über Datenbanksysteme):

- Spreadsheets skalieren nicht besonders gut, d.h. sie haben Schwierigkeiten bei sehr grossen Datenmengen.
- Die Qualität der Daten hängt stark von der Selbstdisziplin der Nutzer ab.
- Es gibt keine Mehrbenutzersynchronisation. Wenn mehrere verschiedene Versionen eines Spreadsheets im Umlauf sind, hat man schnell ein Durcheinander.

5.4 Lektion 4: Datenbanksysteme

Anwendungsentwickler in den 60ern und 70ern haben irgendwann gemerkt, dass sie immer wieder die gleichen Probleme lösen. Einige Entwickler überlegten sich, ein System zu bauen, das diese wiederkehrenden Probleme löst. Auf diese Weise muss nicht jeder Entwickler das Rad neu erfinden. Das senkt die Kosten und verkürzt die Entwicklungszeit von Anwendungen. Damit war die Idee von Datenbanksystemen geboren. Was macht ein Datenbanksystem für uns? Im Folgenden sind die wichtigsten Punkte aufgeführt:

- Erst einmal kann komfortabler auf die Daten zugegriffen werden. Ich muss nicht wissen wie und wo die Daten konkret abgespeichert sind. Ein Datenbanksystem schirmt uns vor den technischen Details ab. Ich kann ja auch Auto fahren, ohne genau zu wissen, wie der Motor funktioniert. Dieses Konzept nennt sich Datenunabhängigkeit.
- Datenbanksysteme stellen auch Anfragesprachen zur Verfügung, die mir helfen, auf die Daten zuzugreifen. Oft sind diese Sprachen deklarativ, d.h. ich sage dem System welche Daten ich brauche. Ich muss nicht wissen, wo und wie diese Daten geholt werden müssen. Darum kümmert sich das System.
- Ausserdem beinhaltet ein Datenbanksystem Verfahren für die Fehlerbehandlung. Wenn ein Datenbanksystem abstürzt, gehen keine Daten verloren. Es wird ständig im Hintergrund mitprotokolliert, was gerade gemacht wird. Dies passiert automatisch, ich muss mich als Benutzer nicht darum kümmern.
- Wenn mehrere Benutzer gleichzeitig Daten verändern kann das zu inkonsistenten Zuständen führen. Wenn z.B. zwei Benutzer eine Datei gleichzeitig öffnen, Änderungen machen und dann die geänderte Datei zurückschreiben, überleben nur die Änderungen die als letztes zurückgeschrieben werden. Ein Datenbanksystem synchronisiert die Zugriffe der verschiedenen Benutzer automatisch und vermeidet inkonsistente Zustände.
- Auch kann ich angeben, wer welche Daten sehen oder gar ändern darf. Ich kann sogar Regeln angeben, auf welche Weise bestimmte Daten geändert werden dürfen. Das Datenbanksystem überwacht auf diese Weise die Integrität der Daten.
- Datenbanksysteme können mit grossen Datenmengen effizient umgehen. Datenbanksysteme können die Daten indexieren, um so schnell

die gewünschte Information zu finden. Dies ist vergleichbar mit einem Index in einem Buch in dem ich Schlagworte und deren Seitennummer nachschlagen kann.

Zusammenfassend zur Funktionalität von Datenbanksystemen lässt sich sagen, dass sie viele der Probleme von Spreadsheets vermeiden.

Datenbanksysteme werden in den verschiedensten Anwendungen eingesetzt, laut Gartner hat der weltweite Markt für Datenbanksysteme eine Grösse von 46 Milliarden US\$. Datenbanksysteme sind oft im Hintergrund tätig, aber fast überall dort zu finden, wo grössere Mengen von Daten systematisch verwaltet werden müssen:

- Bei Banken für die Kontoverwaltung und Transaktionsabwicklung
- Auf Webseiten und Services, wie z.B. Social Networks
- Bei Transportunternehmen
- Selbst auf Smartphones verwenden viele Apps eine Datenbank im Hintergrund

5.5 Lektion 5: Datenmodellierung

Die im Moment am gebräuchlichste Art von Datenbanken sind relationale Datenbanksysteme. Sie speichern die Daten gut strukturiert und sehr systematisch in Tabelle mit Zeilen und Spalten ab. Dabei werden oft auch mathematische Begriffe aus der Relationentheorie verwendet:

- Tupel = Zeile
- Attribut = Spalte
- Tabelle = Relation

Ein weiterer wichtiger Punkt ist, dass eine Relation aus zwei Teilen besteht:

- einem Schema R , das sind die Metadaten
- und einer Instanz R , das sind die eigentlichen Daten, also der Inhalt der Relation

Das folgende Diagramm veranschaulicht das Ganze mit dem SBB-Datensatz. Das Ganze sieht zwar ähnlich wie ein Spreadsheet aus, aber ein entscheidender Unterschied ist, dass das Datenbanksystem die Einhaltung des Schemas durchsetzt.

Schema

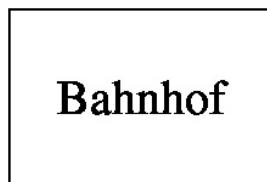
	code character (4)	bahnhof character varying (40)	bezugsjahr integer	kanton character (3)	dtv integer	dwv integer	dnmv integer
1	AD	Aadorf	2018	TG	1700	2000	1000
2	AA	Aarau	2018	AG	37900	44800	22700
3	ABO	Aarburg-Oftringen	2018	AG	2500	3000	1300
4	AAT	Aathal	2018	ZH	740	800	610
5	ACLA	Acla da Fontauna	2018	GR	90	90	80
6	AE	Aesch	2018	BL	2000	2400	1300
7	AF	Affoltern am Albis	2018	ZH	6800	8100	4000
8	AIG	Aigle	2018	VD	6800	7700	5000
9	AI	Airolo	2018	TI	420	370	510
10	ALL	Allaman	2018	VD	3400	4100	1700

Instanz

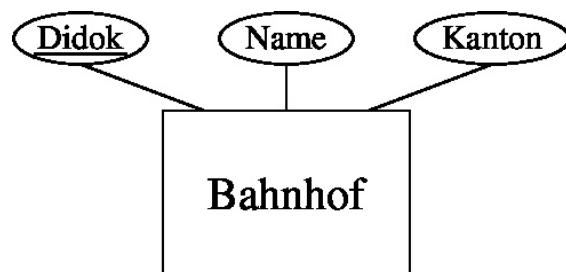
Bevor man anfängt eine Relation anzulegen und mit Daten zu füllen, sollte man sich zuerst einmal Gedanken über die Struktur der Daten machen. Im ersten Schritt entwirft man ein sogenanntes konzeptuelles Schema, was dann in eine oder mehrere Relationen umgesetzt wird. Ein wichtiges konzeptuelles Schema ist das Entity-Relationship-Modell, oder kurz ER-Modell. Dieses Modell stellt grafische Komponenten zur Modellierung zur Verfügung. Das macht es einfacher den Überblick zu wahren und sich mit den Anwendungsexperten auszutauschen. Die gebräuchlichsten Komponenten im ER-Modell sind:

- Entitäten (Entities)
- Attribute
- Schlüssel
- Beziehungen (Relationships)

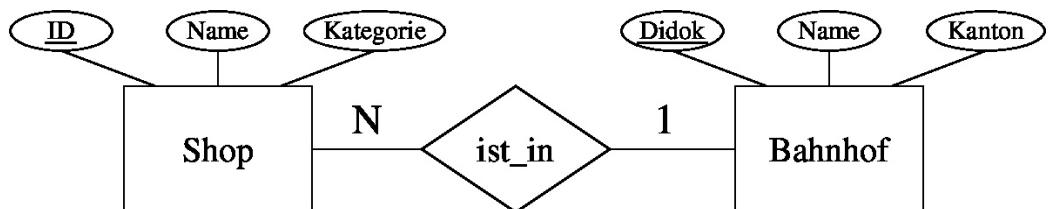
Wir spielen das Ganze nun mit den SBB-Daten durch. Eine Entität ist ein Ding, eine Person, oder ein Konzept. Eine Sammlung von gleichen Entitäten bildet eine (Entitäts-)Menge, die durch ein Rechteck dargestellt wird:



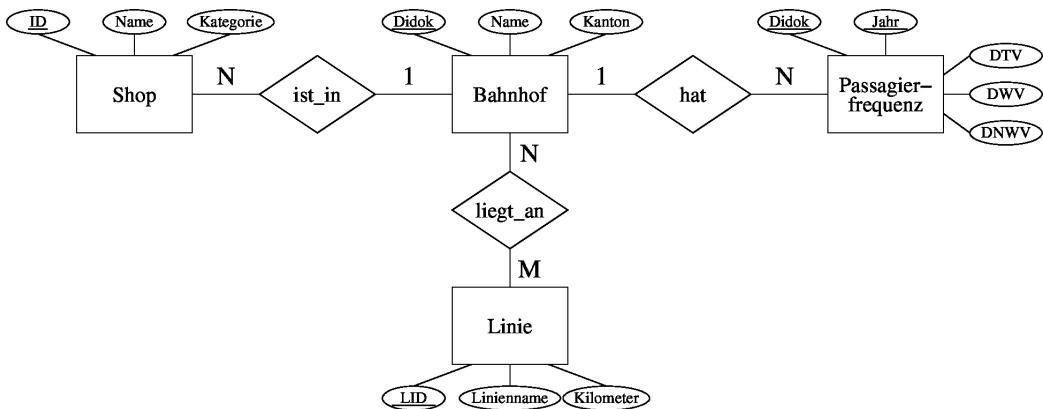
Eigenschaften von Entitäten werden mit Attribut en beschrieben. Ein Attribut wird durch eine Ellipse repräsentiert. Attribute, die eine Entität eindeutig beschreiben, sind Schlüssel und werden im Diagramm unterstrichen:



Eine Beziehung verbindet zwei oder mehr Entitätsmengen und wird mit Hilfe einer Raute dargestellt. Ein Beispiel für eine binäre Beziehung ist die Beziehung zwischen Shop und Bahnhof. Die Symbole N und 1 stehen für Funktionalitäten und sagen aus, dass in einem Bahnhof N, also beliebig viele, Shops sein können, ein (konkreter) Shop aber nur in einem Bahnhof ist:



Das gesamte Modell umfasst noch Linien und Passagierfrequenzen, die über die Beziehungen "hat" und "liegt_an" mit der Entitätsmenge Bahnhof verbunden sind:



Die Umsetzung in Relationen wird in zwei Schritten vorgenommen. Zuerst werden die Entitäten umgesetzt und danach die Beziehungen. Jede Entität wird dabei in eine eigene Relation übersetzt (die jeweiligen Schlüssel sind dabei unterstrichen):

- Bahnhof(Didok, Name, Kanton)
- Shop(ID, Name, Kategorie)
- Passagierfrequenz(Didok, Jahr, DTV, DWV, DNWV)
- Linie(LID, Linienname, Kilometer)

Bei den Beziehungen wird zwischen 1:N und N:M Beziehungen unterschieden. Bei den 1:N Beziehungen wird der Schlüssel der Relation auf die verwiesen wird (die Entität auf der 1-Seite) als Fremdschlüssel in die Relation übernommen (Fremdschlüssel sind kursiv dargestellt):

- Shop(ID, Name, Kategorie, Didok)
- Passagierfrequenz(Didok, Jahr, DTV, DWV, DNWV)
- Die N:M Beziehungen bilden eine eigene Relation:
- Bahnhoflinien(Didok, LID)

5.6 Lektion 6: Datenbankanfragen

SQL (steht für Structured Query Language, auf deutsch: strukturierte Anfrage-sprache) ist eine standardisierte Anfragesprache für relationale Datenbanksysteme. Das bedeutet, dass, von kleineren Ausnahmen abgesehen, alle relationalen Systeme die gleiche Sprache verwenden. SQL ist eine deklarative Sprache, d.h. als Benutzer gebe ich nur an welchen Daten ich interessiert bin, das Datenbanksystem kümmert sich darum, wie die Anfrage verarbeitet wird.

Das Kernstück einer Anfrage ist ein Select-From-Where-Block, in dem die Details der Anfrage spezifiziert werden:

```
select Attributliste  
from Relationenliste  
where Filter;
```

Dabei spielen die drei Teile folgenden Rollen:

- Die Attributliste in der Select-Klausel gibt an, welche Attribute im Ergebnis ausgegeben werden.
- Die Relationenliste in der From-Klausel gibt an, aus welchen Relationen die Daten kommen.
- Mit dem Filter in der Where-Klausel kann spezifiziert werden, welche Daten mich genau interessieren.

Wir spielen das Ganze einmal mit verschiedenen Anfragen an den SBB-Datensatz durch. Eine sehr einfache Anfrage gibt den gesamten Inhalt einer Relation aus:

```
select *  
from Bahnhof;
```

Der Stern (*) ist eine Abkürzung für “gib alle Attribute aus”. Die obige Anfrage gibt den kompletten Inhalt der Relation Bahnhof aus:

Didok	Name	Kanton
3000	Zürich HB	ZH
23	Liestal	BL
1026	Genève Aéroport	GE
...

Anstatt alle Attribute auszugeben, können wir uns auch auf bestimmte Attribute beschränken:

```
select Name, Kanton
from Bahnhof;
```

Im Ergebnis dieser Anfrage fehlt nun die Spalte Didok:

Name	Kanton
Zürich HB	ZH
Liestal	BL
Genève Aéroport	GE
...	...

Die Where-Klausel ist der interessanteste Teil einer Anfrage. Hier können wir z.B. ungewollte Tupel ausfiltern:

```
select *
from Bahnhof
where Kanton = "ZH";
```

Hier sind jetzt wieder alle Spalten dabei, allerdings werden nur die Tupel der Bahnhöfe im Kanton Zürich in das Ergebnis übernommen:

Didok	Name	Kanton
3000	Zürich HB	ZH
3006	Zürich Oerlikon	ZH
...

Bisher haben wir nur Anfragen betrachtet die sich auf eine Relation beziehen. Es kann aber sein, dass die benötigten Daten über mehrere Tabellen verteilt sind. Wir müssen spezifizieren wie Tupel aus verschiedenen Tabellen miteinander verknüpft werden. Auch das geschieht in der Where-Klausel. In der nächsten Anfrage verknüpfen wir die Passagierfrequenzdaten mit den Namen der Bahnhöfe an denen die Statistiken erhoben wurden:

```
select B.Name, P.Jahr, P.DTV
from Bahnhof B, Passagierfrequenz P
where B.Didok = P.Didok;
```

Wir sagen dem Datenbanksystem, dass Tupel aus der Relation Bahnhof nur mit Tupeln aus der Relation Passagierfrequenz verbunden werden dürfen, wenn ihr Wert für Didok übereinstimmt. Die Abkürzungen B und P für die Relationen wurden eingeführt um sich Tipparbeit zu sparen (ansonsten müsste man an allen Stellen an denen B und P auftauchen, den vollen Namen angeben). Das Ergebnis dieser Anfrage sieht folgendermassen aus:

B.Name	P.Jahr	P.DTV
Zürich HB	2018	423600
Liestal	2017	17500
...

Die Filter in einer Where-Klausel können beliebig komplex werden. So kann man das Ergebnis der Verknüpfung der Daten aus den beiden Tabellen Bahnhof und Passagierfrequenz weiter einschränken. In der nächsten Anfrage sind wir nur an den Daten aus dem Kanton Zürich für das Jahr 2018 interessiert:

```
select B.Name, P.Jahr, P.DTV
from Bahnhof B, Passagierfrequenz P
```

```
where B.Didok = P.Didok
and P.Jahr = 2018
and B.Kanton = "ZH";
```

Die einzelnen Filterausdrücke werden mit dem Schlüsselwort "and" verknüpft. Es ist auch möglich, Bedingungen mit den logischen Operatoren "or" oder "not" zu formulieren. Für diese Anfrage sähe das Ergebnis folgendermassen aus:

B.Name	P.Jahr	P.DTV
Zürich HB	2018	423600
Zürich Oerlikon	2018	80000
...

Das war jetzt nur ein kurzer Überblick über die Funktionalität von SQL. Es ist noch viel mehr möglich, z.B. das Gruppieren von Daten und das Erstellen einfacher Statistiken. Als einfaches Beispiel wird hier noch das Sortieren von Daten mit Hilfe des Schlüsselworts "order by" gezeigt:

```
select B.Name, P.Jahr, P.DTV
from Bahnhof B, Passagierfrequenz P
where B.Didok = P.Didok
and P.Jahr = 2018
order by P.DTV desc;
```

"desc" steht für descending, auf deutsch absteigend. Man kann auch aufsteigend sortieren: dies geschieht mit dem Schlüsselwort "asc". Wenn man weder "asc" noch "desc" angibt, wird automatisch aufsteigend sortiert. Hier noch das Ergebnis der Beispielanfrage:

B.Name	P.Jahr	P.DTV
Zürich HB	2018	423600
Bern	2018	184000
...

Bisher haben wir nur Anfragen betrachtet. Bevor wir aber Anfragen an ein Datenbanksystem stellen können, müssen wir zuerst Relationen anlegen und mit Inhalt füllen. Eventuell wollen wir später auch wieder Daten ändern oder sogar löschen. Auch das ist mit SQL möglich.

Das Schema einer Relation wird mit dem Kommando “create table” angelegt. Um z.B. die Relation für Bahnhöfe anzulegen, wird folgender Befehl abgesetzt:

```
create table Bahnhof (
    Didok integer,
    Name varchar(80),
    Kanton char(2),
    primary key (Didok)
)
```

Dabei wird für jedes Attribut angegeben, welchen Datentyp es hat: Ganzzahlen für Didok, eine Zeichenkette variabler Länge (maximal 80 Zeichen) für Name und eine Zeichenkette der Länge 2 für Kanton. Es kann ausserdem angegeben werden, welches Attribut den Schlüssel bildet.

Um Daten einzufügen wird der Befehl “insert” verwendet. Man gibt den Namen der Relation in die man etwas einfügen möchte an, gefolgt von den Werten:

```
insert into Bahnhof
values (3000, 'Zürich', 'ZH');
```

Um Daten wieder zu löschen verwendet man den Befehl “delete”. Dieser Befehl hat, ähnlich wie eine Anfrage, eine From- und eine Where-Klausel, mit denen man angibt welche Daten gelöscht werden sollen, hier z.B. den Bahnhof mit Didok 3000:

```
delete from Bahnhof
where Didok = 3000;
```

Man sollte beim Löschen allerdings vorsichtig sein. Bei einer leeren Where-Klausel wird der gesamte Inhalt der Relation gelöscht. Wenn man sich nicht ganz sicher ist, kann man die From- und Where-Klausel eines Delete-Befehls

einfach vorher mit einer Select-Klausel ausführen. Dann sieht man, welche Daten gelöscht würden.

Zu guter Letzt können Daten in einer Relation auch geändert werden. Der folgende Befehl weist dem Attribut Name des Tupels mit der Didok 3000 den Wert "Zürich HB" zu:

```
update Bahnhof  
set Name = "Zürich HB"  
where Didok = 3000;
```

5.7 Lektion 7: Big Data und die Cloud

Ein Laptop- oder Desktopcomputer ist gut geeignet um Spreadsheets zu erstellen und zu analysieren, um einen Text zu editieren oder um einen Film zu schauen. Wenn die Datenmengen allerdings etwas grösser werden, stösst man schnell an die Grenzen. Um mit grossen Datenmengen, Schlagwort "Big Data", zurechtzukommen gibt es im Prinzip zwei Möglichkeiten:

- Scale-Up: man kauft sich einen grösseren Rechner mit mehr Speicherplatz und einem stärkeren Prozessor. Das wird auf deutsch auch gelegentlich als vertikale Skalierung bezeichnet.
- Scale-Out: hier verteilt man die Arbeit auf viele (kleinere) Maschinen. Oft kommen handelsübliche PCs mit Linux zum Einsatz. Diese Methode wird auch horizontale Skalierung genannt.

Der Vorteil von Scale-Up ist, dass kaum etwas geändert werden muss. Im Prinzip läuft die gleiche Software einfach auf einem grössseren Rechner. Ein Nachteil ist, dass dies nicht beliebig weit getrieben werden kann. Irgendwann hat man die grösste Ausbaustufe erreicht. Beim Scale-Out gibt es (theoretisch) keine Grenzen: wenn man mehr Kapazitäten benötigt, kauft man einfach Geräte dazu. So gibt es Datenzentren mit zehn- oder sogar hunderttausenden von Maschinen. Ein Nachteil ist, dass man die Software anpassen und parallelisieren muss: die Rechenprozesse auf den verschiedenen Computern müssen synchronisiert werden, was bei vielen Rechnern zu einem erhöhten Kommunikationsaufwand führt. Außerdem benötigt man ausgefeilte Mechanismen zur Ausfallsicherheit: bei so vielen Maschinen geht eigentlich ständig etwas kaputt.

Anfang der 2000er-Jahre entwickelten Firmen wie Google und Yahoo Systeme, wie z.B. das Google File System, Map-Reduce und den Vorläufer von Hadoop,

um mit extrem grossen Datenmengen zurechtzukommen. Viele dieser System, wie z.B. Hadoop, gibt es inzwischen in einer frei erhältlichen Open-Source-Variante. Inzwischen gibt es auch Systeme die einfacher und komfortabler zu bedienen sind, wie z.B. Apache Spark. Auch Datenbanksysteme durchlaufen im Moment einige Änderungen, mit den sogenannten NoSQL-Systemen wird versucht Datenbanksysteme zu entwickeln, die für einen Scale-Out geeignet sind.

Auch wenn man selbst kein Rechenzentrum zuhause hat, kann man Datenanalysen auf grossen Datenmengen durchführen. Viele Cloudanbieter haben Plattformen entwickelt, auf denen man sich Ressourcen mieten kann. Beispiele sind AWS (Amazon Cloud Service) und EC2 (Elastic Compute Cloud) von Amazon, DataProc von Google und Azure von Microsoft.

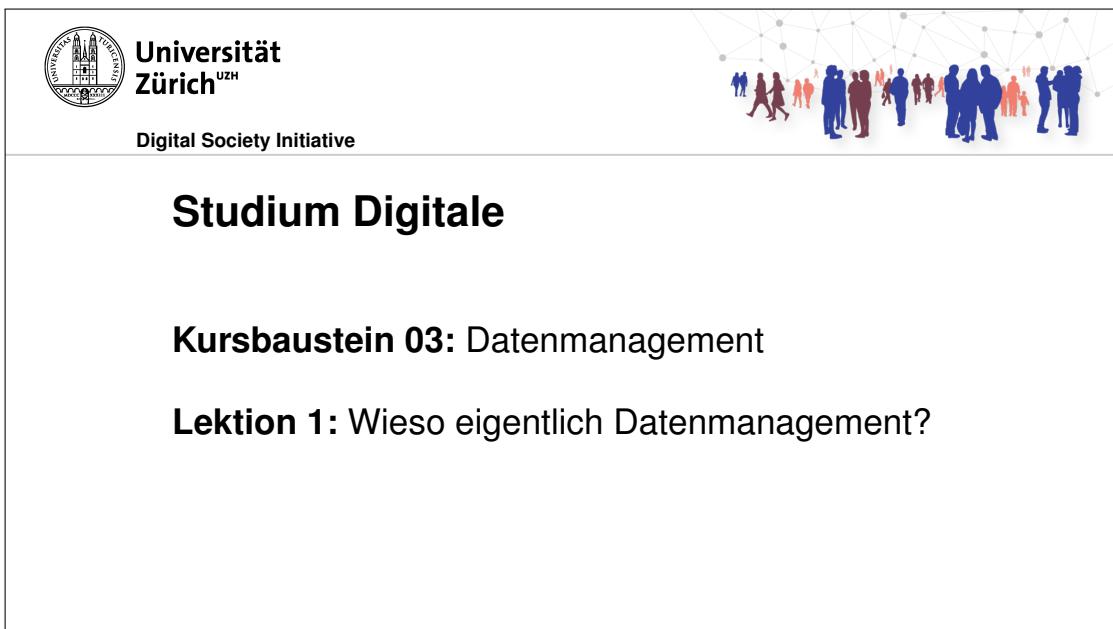
5.8 Lektion 8: Rückblick und Ausblick

In diesem Kursbaustein haben Sie einen Überblick über den Themenbereich Datenmanagement erhalten und Sie wissen jetzt:

- mit welchen Datenmengen wir heutzutage konfrontiert werden und dass es wichtig ist, mit diesen Daten sorgfältig umzugehen,
- welchen Lebenszyklus Daten durchlaufen,
- was Spreadsheets sind und wie sie eingesetzt werden,
- welche Vorteile Datenbanksysteme haben,
- wie Daten modelliert werden,
- wie Daten in einem Datenbanksystem abgefragt werden
- und was es mit Big-Data auf sich hat.

Es gibt noch viel mehr zu erzählen, aber wir haben nur begrenzt Zeit. Eventuell hören Sie im Rahmen Ihres Studiums noch weitere, tiefergehende Vorlesungen zu diesem Thema. Auch wenn im Moment nicht klar ist, wie zukünftige Technologien in diesem Bereich aussehen werden, eins ist sicher: das Thema Datenmanagement wird noch eine ganze Weile relevant sein.

5.9 Folien



The image shows a slide from the University of Zurich's Digital Society Initiative. At the top left is the University of Zurich logo, which is a circular emblem featuring a building and the text "UNIVERSITÄT ZÜRICH". To its right, the text "Universität Zürich" is written in a bold, black, sans-serif font, with "UZH" in a smaller, superscript-like font. Below this, the text "Digital Society Initiative" is printed in a smaller, regular black font. To the right of the text is a graphic of several stylized human figures in red and blue, standing in a line and connected by a network of grey dots, symbolizing connectivity and digital society.

Studium Digitale

Kursbaustein 03: Datenmanagement

Lektion 1: Wieso eigentlich Datenmanagement?

Produzierte Daten pro Tag

- 500 Mill. Tweets
- 294 Mrd. E-Mails
- Facebook: 4PByte Daten ~ 850'000 DVDs

(Raconteur 2019)

Schätzungen für 2025

- 463 Exabytes pro Tag
- knapp 100 Mrd. DVDs

Um was geht es bei Datenmanagement?

Kurz gesagt:

- Es geht um das Sammeln, Speichern und Verarbeiten von Daten
- Das Ganze sollte zuverlässig, sicher und effizient geschehen
- Macht auch schon Sinn für weniger als Millionen von DVDs...

Was geht mich das an?

- Forschungsprojekte
- Industrie 4.0

Was gibt es zu beachten?

- Wie finde ich, was ich suche?
- Hilfe, mein Rechner stürzt ab!

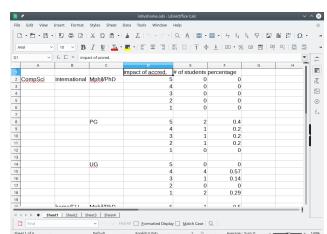
Dateiformate

- Textdateien
- Worddateien
- Spreadsheets
- PDFs
- JSON
- XML
- HTML
- ...

Was mache ich hiermit?

2	5.3	$\frac{1}{2}$	z	δ
3	4.6	—	a	γ
2	10.2	$\frac{3}{4}$	r	δ
...

Mehrere Benutzer



Sicherheit

- Datenschutz
- Privatsphäre

Übersicht über diesen Kursbaustein:

- Was gibt es zu beachten?
- Lebenszyklus von Daten
- Professionelle Werkzeuge
- Datenmodellierung
- Datenbankanfragen
- Big Data und die Cloud

Referenzen

Raconteur (2019). *A Day in Data*.

<https://www.raconteur.net/infographics/a-day-in-data>.

Referent

Dr. Sven Helmer
Institut für Informatik

© Universität Zürich
Digital Society Initiative



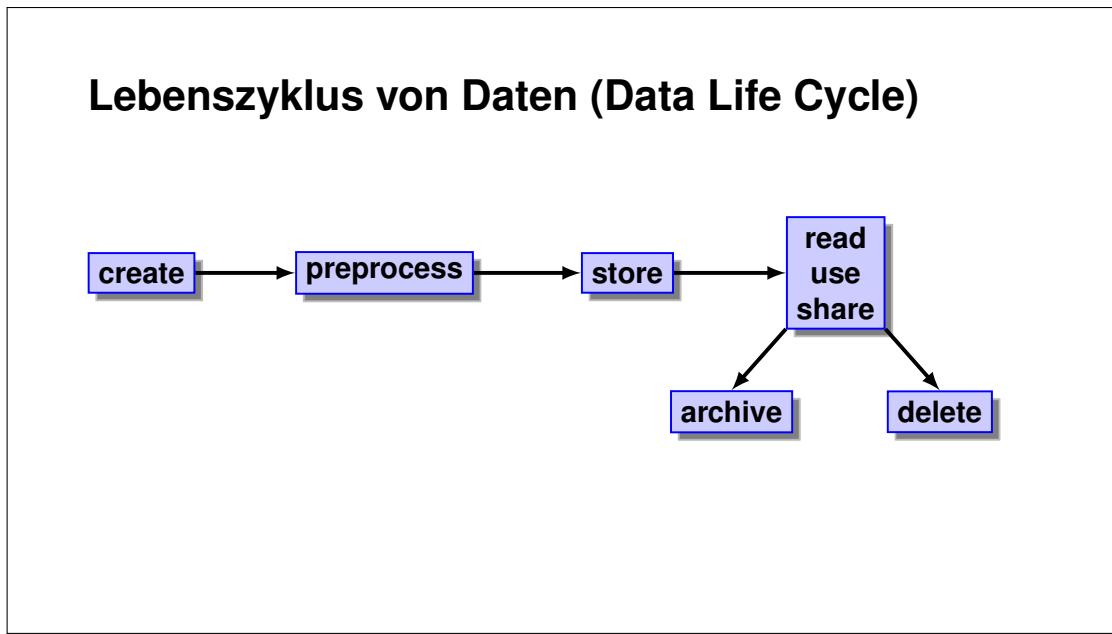
Universität
Zürich UZH

Digital Society Initiative

Studium Digitale

Kursbaustein 03: Datenmanagement

Lektion 2: Lebenszyklus von Daten



1. Daten erzeugen

- Digitalisierung
- Logdateien
- Sensoren

2. Daten vorverarbeiten

- Qualitätsprobleme

3. Daten speichern

- müssen auffindbar sein
- und möglichst schnell

4. Daten nutzen**5. Archivierung/Lösung**

- keine aktive Verwendung
- rechtliche Gründe

Viele Daten leben lange

- Wegen Langlebigkeit macht sorgfältiges Datenmanagement Sinn
- Daten können für andere Leute interessant sein:
 - SBB-Daten: <https://data.sbb.ch>
 - Open-Data-Initiativen: <https://opendata.swiss>
 - Interessante Kombinationen

Referent

Dr. Sven Helmer
Institut für Informatik

© Universität Zürich
Digital Society Initiative



Universität
Zürich^{UZH}

Digital Society Initiative



Studium Digitale

Kursbaustein 03: Datenmanagement

Lektion 3: Spreadsheets

Erster Spreadsheet

VisiCalc (siehe auch Grad 2007)

- Eine Idee Ende 1978
- von D. Bricklin und B. Frankston
- Vorläufer von Excel

Was ist eigentlich ein Spreadsheet?

Kurz gesagt: ein Werkzeug für die

- Eingabe,
- Speicherung,
- Analyse und
- Visualisierung

von Daten.

Aufbau eines Spreadsheets

Ein Spreadsheet besteht aus einer Tabelle mit Zeilen und Spalten:

- oberste Zeile: Beschreibungen
- jede Zeile darunter: ein Datensatz
- jede Spalte: eine Variable
- jede Zelle: ein Wert (kann auch berechnet werden)

Ein Beispiel

	A	B	C	D	E	F	G	H
1	Code	Bahnhof	Haltestelle	Bezugsjahr	Kanton	DTV	DWV	DNWV
2	AD	Aadorf		2018	TG	1700	2000	1000
3	AA	Aarau		2018	AG	37900	44800	22700
4	ABO	Aarburg-Ottringen		2018	AG	2500	3000	1300
5	AAT	Aathal		2018	ZH	740	800	610
6	ACLA	Acia da Fontauna		2018	GR	90	90	80
7	AE	Aesch		2018	BL	2000	2400	1300
8	AF	Affoltern am Albis		2018	ZH	6800	8100	4000
9	AIG	Aigle		2018	VD	6800	7700	5000
10	AI	Airolo		2018	TI	420	370	510

Generelle Best Practices

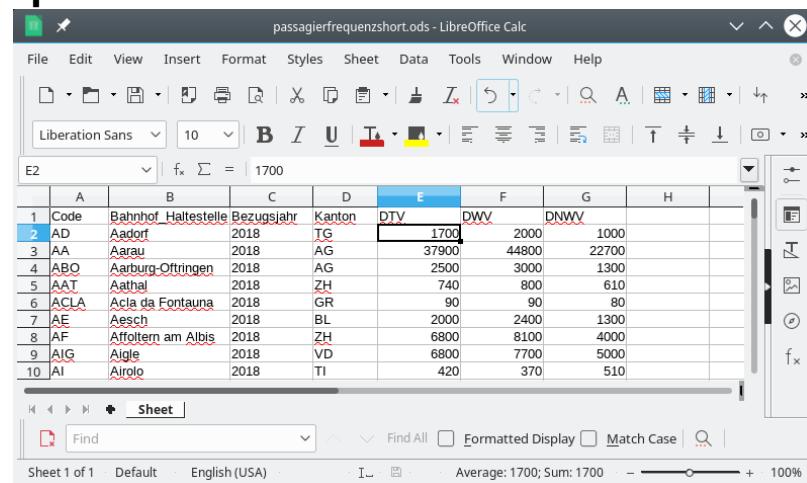
- Namensgebung
- Metadaten anlegen
- Backups
- Als Text abspeichern

Best Practices für Spreadsheets

- Nur ein Wert pro Zelle
- Zellen nicht einfach leer lassen
- Keine Berechnungen auf den Rohdaten
- Datenvielfältigung benutzen
- Einfach einmal einen Datensatz in einem Satz zusammenfassen

(aus Broman und Woo 2018)

Beispiel



The screenshot shows a LibreOffice Calc spreadsheet titled "passagierfrequenzshort.ods". The data is organized in a table with the following columns: Code, Bahnhof, Haltestelle, Bezugsjahr, Kanton, DTV, DWV, and DNWV. The rows contain data for various locations and years, such as Aadorf, Aarau, Aarburg-Ottringen, Aathal, Acla da Fontana, Aesch, Aftoltern am Albis, Aigle, and Ajrolo. The "DTV" column is highlighted in blue, indicating it is the active cell.

	A	B	C	D	E	F	G	H
1	Code	Bahnhof	Haltestelle	Bezugsjahr	Kanton	DTV	DWV	DNWV
2	AD	Aadorf		2018	TG	1700	2000	1000
3	AA	Aarau		2018	AG	37900	44800	22700
4	ABO	Aarburg-Ottringen		2018	AG	2500	3000	1300
5	AAT	Aathal		2018	ZH	740	800	610
6	ACLA	Acla da Fontana		2018	GR	90	90	80
7	AE	Aesch		2018	BL	2000	2400	1300
8	AF	Affoltern am Albis		2018	ZH	6800	8100	4000
9	AIG	Aigle		2018	VD	6800	7700	5000
10	AI	Ajrolo		2018	TI	420	370	510

Vorteile und Grenzen von Spreadsheets

- + Es ist einfach Daten einzutragen und loszulegen
- + Daten können auch importiert und exportiert werden
- Bei Millionen von Datensätzen wird es mühsam
- Probleme mit der Datenqualität
- Keine Mehrbenutzersynchronisation

Referenzen

- Broman, Karl W. und Kara H. Woo (2018). "Data Organization in Spreadsheets". In: *The American Statistician* 72.1, S. 2–10.
- Grad, B. (2007). "The Creation and the Demise of VisiCalc". In: *IEEE Annals of the History of Computing* 29.3, S. 20–31.

Referent

Dr. Sven Helmer
Institut für Informatik

© Universität Zürich
Digital Society Initiative



Universität
Zürich^{UZH}

Digital Society Initiative



Studium Digitale

Kursbaustein 03: Datenmanagement

Lektion 4: Datenbanksysteme

Ursprung von Datenbanksystemen

60er und 70er: Immer wieder die gleichen Probleme

Warum kein System bauen, das diese Probleme löst?

- Senkt Kosten und Entwicklungszeit
- Die Idee von Datenbanksystemen war geboren

Was macht ein Datenbanksystem für uns?

- Datenunabhängigkeit
- Anfragesprachen
- Fehlerbehandlung
- Mehrbenutzersynchronisation
- Datenintegrität
- Skalierbarkeit

(siehe auch Kemper und Eickler 2015)

Bekannte Systeme

- Oracle
- Microsoft SQLServer
- IBM DB2
- MySQL
- PostgreSQL

Wo stecken diese ganzen Systeme?

- Banken und andere Finanzunternehmen
- Verkehrsunternehmen
- Im Backend von Apps
- eigentlich überall dort wo Daten verarbeitet werden
- Datenbanken sind eine Selbstverständlichkeit geworden

Referenzen

Kemper, Alfons und André Eickler (2015). *Datenbanksysteme – Eine Einführung*. 10. Aufl. Walter de Gruyter.

Referent

Dr. Sven Helmer
Institut für Informatik

© Universität Zürich
Digital Society Initiative



Universität
Zürich^{UZH}

Digital Society Initiative



Studium Digitale

Kursbaustein 03: Datenmanagement

Lektion 5: Datenmodellierung

Relationale Datenbanken

In einem relationalen Datenbanksystem

- werden die Daten in Tabellen abgespeichert.
- Die Tabellen bestehen aus Zeilen und Spalten.

Es werden oft auch mathematische Begriffe verwendet:

- Tupel = Zeile
- Attribut = Spalte
- Relation = Tabelle

(siehe auch Kemper und Eickler 2015)

Beispiel (SBB):

	code character (4)	bahnhof character varying (40)	bezugsjahr integer	kanton character (3)	dtv integer	dwv integer	dnmv integer
1	AD	Aadorf	2018	TG	1700	2000	1000
2	AA	Aarau	2018	AG	37900	44800	22700
3	ABO	Aarburg-Oftringen	2018	AG	2500	3000	1300
4	AAT	Aathal	2018	ZH	740	800	610
5	ACLA	Acla da Fontauna	2018	GR	90	90	80
6	AE	Aesch	2018	BL	2000	2400	1300
7	AF	Affoltern am Albis	2018	ZH	6800	8100	4000
8	AIG	Aigle	2018	VD	6800	7700	5000
9	AI	Airolo	2018	TI	420	370	510
10	ALL	Allaman	2018	VD	3400	4100	1700

Aufbau von Relationen

Eine Relation besteht aus zwei Teilen:

- einem Schema \mathcal{R} , das sind die Metadaten
- einer Instanz R , das ist der Inhalt der Relation

Beispiel:

Schema

	code character (4)	bahnhof character varying (40)	bezugsjahr integer	kanton character (3)	dtv integer	dwv integer	dnmv integer
1	AD	Aadorf	2018	TG	1700	2000	1000
2	AA	Aarau	2018	AG	37900	44800	22700
3	ABO	Aarburg-Oftringen	2018	AG	2500	3000	1300
4	AAT	Aathal	2018	ZH	740	800	610
5	ACLA	Acla da Fontauna	2018	GR	90	90	80
6	AE	Aesch	2018	BL	2000	2400	1300
7	AF	Affoltern am Albis	2018	ZH	6800	8100	4000
8	AIG	Aigle	2018	VD	6800	7700	5000
9	AI	Airolo	2018	TI	420	370	510
10	ALL	Allaman	2018	VD	3400	4100	1700

Instanz

Datenbankentwurf

- Zuerst konzeptuelles Schema
- Dann Umsetzung in Relationen
- Ein wichtiges konzeptuelles Modell:
 - Entity-Relationship-Modell

Entity-Relationship-Modell

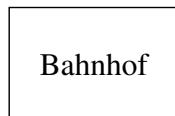
Das Entity-Relationship-Modell hat folgende Grundbausteine:

- Entitäten (Entities)
- Attribute
- Schlüssel
- Beziehungen (Relationships)

Wir spielen das einmal mit dem SBB-Beispiel durch

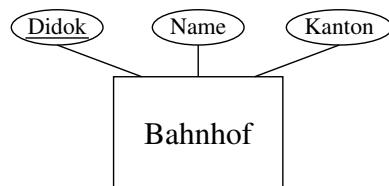
Entitäten

- Eine *Entität* ist ein Ding, eine Person, ein Ort oder ein Konzept
- Eine Sammlung von gleichen Entitäten bildet eine *Menge*
- Entitätsmengen werden durch Rechtecke dargestellt:



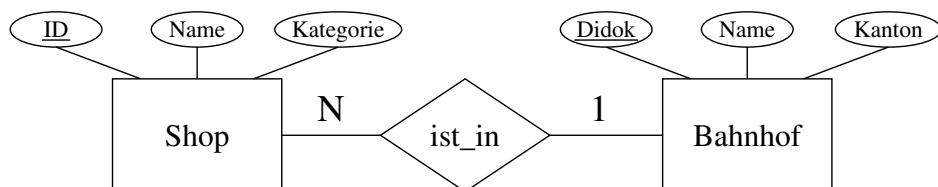
Attribute

- Eigenschaften von Entitäten werden mit *Attributen* beschrieben
- Ein Attribut, das eine Entität eindeutig beschreibt, ist ein *Schlüssel*
- Attribute werden durch Ellipsen repräsentiert:

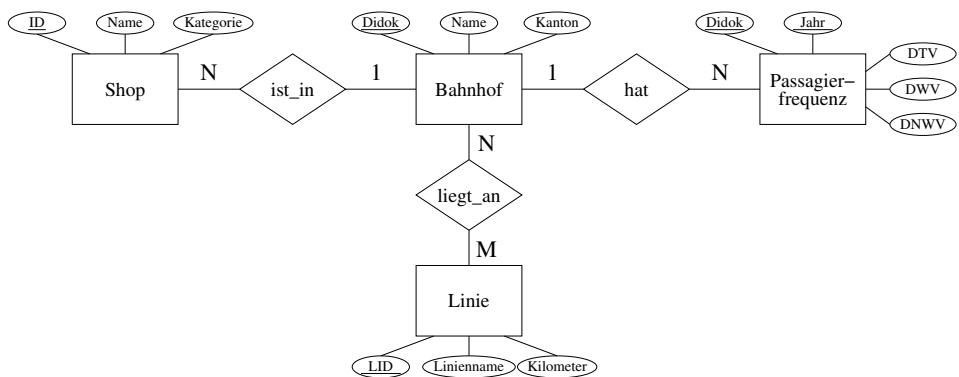


Beziehungen

- Eine *Beziehung* verbindet zwei oder mehr Entitätsmengen
- Darstellung durch eine Raute mit Funktionalitäten
- Beispiel für eine binäre Beziehung



Und nun alles zusammen



Umsetzung in Tabellen

- Bahnhof (Didok, Name, Kanton)
- Shop (ID, Name, Kategorie)
- Passagierfrequenz (Didok, Jahr, DTV, DWV, DNWV)
- Linie (LID, Liniename, Kilometer)

Was machen wir mit den Beziehungen?

Beziehungen

1:N

- Shop (ID, Name, Kategorie, *Didok*)
- Passagierfrequenz (*Didok*, Jahr, DTV, DWV, DNWV)

N:M

- Bahnhoflinien (*Didok*, LID)

Und nun alles zusammen

- Bahnhof (Didok, Name, Kanton)
- Shop (ID, Name, Kategorie, *Didok*)
- Passagierfrequenz (*Didok*, Jahr, DTV, DWV, DNWV)
- Linie (LID, Liniename, Kilometer)
- Bahnhoflinien (*Didok*, LID)

Referenzen

Kemper, Alfons und André Eickler (2015). *Datenbanksysteme – Eine Einführung*. 10. Aufl. Walter de Gruyter.

Referent

Dr. Sven Helmer
Institut für Informatik

© Universität Zürich
Digital Society Initiative



Universität
Zürich^{UZH}

Digital Society Initiative



Studium Digitale

Kursbaustein 03: Datenmanagement

Lektion 6: Datenbankanfragen

SQL

- SQL steht für Structured Query Language
- und ist **die** Anfragesprache für relationale Datenbanksysteme

SQL-Anfragen

SQL-Anfragen bestehen im Wesentlichen aus drei Teilen:
Select, From, und Where

select	Attributliste
from	Relationenliste
where	Filter;

Ein einfaches Beispiel

Wir möchten die gesamte Bahnhofrelation ausgeben:

```
select      *
from       Bahnhof;
```

Didok	Name	Kanton
3000	Zürich HB	ZH
23	Liestal	BL
1026	Genève-Aéroport	GE
...

Attribute selektieren

Wir können uns auch auf bestimmte Attribute beschränken:

```
select      Name, Kanton  
from       Bahnhof;
```

Name	Kanton
Zürich HB	ZH
Liestal	BL
Genève-Aéroport	GE
...	...

Die Where-Klausel

Mit der Where-Klausel können wir Tupel ausfiltern:

```
select      *  
from       Bahnhof  
where      Kanton = "ZH";
```

Didok	Name	Kanton
3000	Zürich HB	ZH
3006	Zürich Oerlikon	ZH
...

Relationen verbinden

Informationen können über mehrere Relationen verteilt sein.
Dann müssen wir die Informationen zusammensetzen:

```
select      B.Name, P.Jahr, P.DTV
from        Bahnhof B, Passagierfrequenz P
where       B.Didok = P.Didok;
```

B.Name	P.Jahr	P.DTV
Zürich HB	2018	423600
Liestal	2017	17500
...

Kombination mit anderen Filtern

```
select      B.Name, P.Jahr, P.DTV
from        Bahnhof B, Passagierfrequenz P
where       B.Didok = P.Didok
and         P.Jahr = 2018
and         B.Kanton = "ZH";
```

B.Name	P.Jahr	P.DTV
Zürich HB	2018	423600
Zürich Oerlikon	2018	80000
...

Sortieren

```
select      B.Name, P.Jahr, P.DTV
from        Bahnhof B, Passagierfrequenz P
where       B.Didok = P.Didok
and         P.Jahr = 2018
order by    P.DTV desc;
```

B.Name	P.Jahr	P.DTV
Zürich HB	2018	423600
Bern	2018	184000
...

Was geht noch?

Gruppieren und einfache Statistiken, z.B.

- Summe
- Anzahl
- Durchschnitt
- Max/Min

Daten ändern (siehe nächste Folien)

- Relationen anlegen
- Daten einfügen
- Daten löschen
- Daten aktualisieren

Relation anlegen

Dafür müssen wir ihr Schema definieren:

```
create table Bahnhof (
    Didok      integer,
    Name       varchar(80),
    Kanton     char(2),
    primary key (Didok)
);
```

Daten einfügen

Jetzt können wir Daten einfügen:

```
insert into Bahnhof
values (3000, 'Zürich', 'ZH');
```

Daten löschen

Wir können Daten auch wieder loswerden:

```
delete from Bahnhof  
where Didok = 3000;
```

Daten löschen

Wir können Daten auch wieder loswerden:

```
delete from Bahnhof  
where Didok = 3000;
```

Vorsicht: bei leerer Where-Klausel wird alles gelöscht!

Daten aktualisieren

Daten können auch korrigiert/aktualisiert werden:

```
update Bahnhof  
set     Name = 'Zürich HB'  
where   Didok = 3000;
```

Referent

Dr. Sven Helmer
Institut für Informatik

© Universität Zürich
Digital Society Initiative



Universität
Zürich^{UZH}

Digital Society Initiative

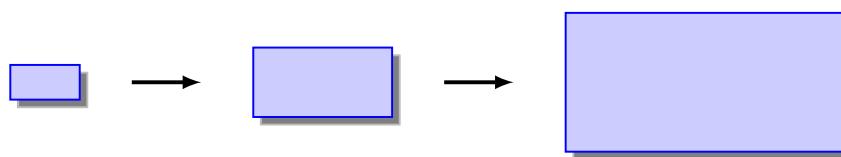


Studium Digitale

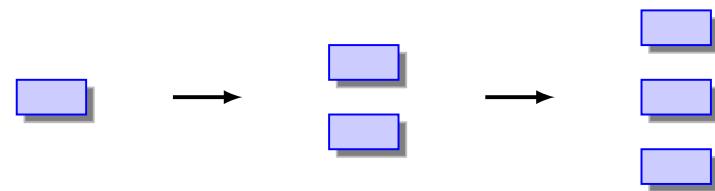
Kursbaustein 03: Datenmanagement

Lektion 7: Big Data und die Cloud

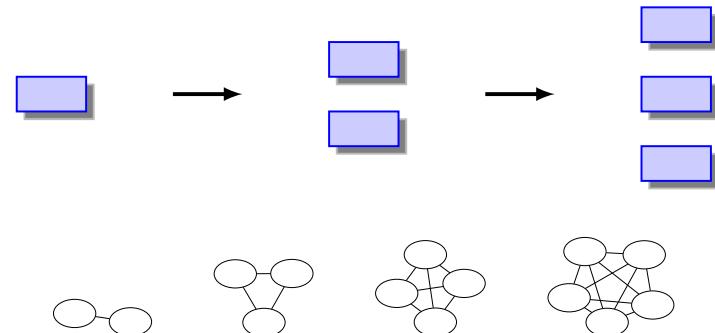
Scale-Up



Scale-Out



Scale-Out



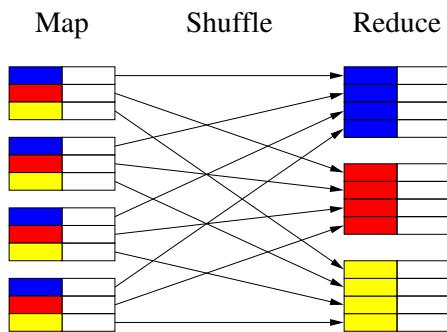
Ende 1990er

- Firmen wie Amazon, Ebay, Google, Yahoo produzieren riesige Datenmengen
- Diese Daten passen nicht mehr auf einen einzelnen Computer

Fehlertolerante Systeme

- Google entwickelt das Google File System und Map-Reduce
- Yahoo: Hadoop Open-Source Framework (samt Dateisystem)
- Apache: Spark
- Parallel zu allem: NoSQL-Datenbanksysteme

Map-Reduce



Beispiel

Map	Shuffle	Reduce
Dokument 1: Heile, heile Segen	(Heile, 1) (heile, 1) (Segen, 1)	a - i: (alles, 1) (heile, 3)
Dokument 2: sieben Tage Regen	(sieben, 1) (Tage, 1) (Regen, 1)	j - s: (Regen, 1) (Segen, 1) (sein, 1)
Dokument 3: sieben Tage Sonnenschein	(sieben, 1) (Tage, 1) (Sonnenschein, 1)	(sieben, 2) (Sonnenschein, 1)
Dokument 4: wird alles wieder heile sein	(wird, 1) (alles, 1) (wieder, 1) (heile, 1) (sein, 1)	t - z: (Tage, 2) (wieder, 1) (wird, 1)

Big-Data Analytics

- Die Verarbeitung von grossen Datenmengen ist heute für fast jeden/jede möglich
- Ressourcen können gemietet werden:
 - Amazon (AWS, EC2)
 - Google Cloud DataProc
 - Microsoft Azure HDInsight
 - (Azure) Databricks

Referent

Dr. Sven Helmer
Institut für Informatik

© Universität Zürich
Digital Society Initiative



Universität
Zürich^{UZH}

Digital Society Initiative



Studium Digitale

Kursbaustein 03: Datenmanagement

Lektion 8: Rückblick und Ausblick

Kurze Geschichte des Datenmanagements

3400 - 3100 v.Chr.(?):	Erfindung der Schrift
2000 v.Chr.:	sumerische Tontafellisten
300 v.Chr.:	Bibliothek von Alexandria
1440:	Erfindung des Buchdrucks
16. Jhd.:	erste Buchindizes

Buchindex

Antilope	Adler	Ader
Ameise	Ader	Adler
Ader	Ameise	Ameise
Ansicht	Ansicht	Ansage
Adler	Antilope	Ansicht
Ansage	Ansage	Antilope

B...

Referent

Dr. Sven Helmer
Institut für Informatik

© Universität Zürich
Digital Society Initiative

6 Digitalisierung

PROF. ABRAHAM BERNSTEIN

6.1 Digitale Daten



The logo of the University of Zurich (UZH) is located on the left side of the slide. It consists of a circular seal with a building and text around it, followed by the text "Universität Zürich UZH". To the right of the logo is a graphic illustration of a network or social structure, showing a group of stylized human figures connected by a web of lines and dots.

Studium Digitale

Kursbaustein 2: Digitalisierung von Information



Universität
Zürich^{UZH}

Digital Society Initiative



Studium Digitale

Kursbaustein 2: Digitalisierung von Information

Lektion 1: Digitale Daten

Prof. Abraham Bernstein
UZH, Institut für Informatik und Digital Society Initiative

Diese Lektion:

- Einleitung
- Digitalisierung von Text
- Digitalisierung von Ton
- Digitalisierung von Bild

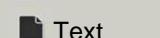
Diese Lektion:

- **Einleitung**
- Digitalisierung von Text
- Digitalisierung von Ton
- Digitalisierung von Bild

Information kann in verschiedenen Ausprägungen vorliegen:



Ton



Bild, Film



Text

Der Mensch nimmt Informationen mit Sinnesorganen auf. Computer benötigen diese in einem Binärformat.

By Jack Cai, CC BY 3.0, softicons.com

Was ist Binär? Von Dezimal- zu Binärzahlen

Computer kennen nur zwei Grundzustände:

Strom fliesst oder kein Strom fliesst. Dies wird als 1 und 0 abstrahiert.

Wir sind gewöhnt, im Dezimalsystem zu rechnen:

...	1000	100	10	1
	10^3	10^2	10^1	10^0

Beispiel

$$\begin{array}{cccc} 2 & 0 & 1 & 3 \\ 2 \times 10^3 & 0 \times 10^2 & 1 \times 10^1 & 3 \times 10^0 \end{array}$$

Von Dezimal- zu Binärzahlen

Computer benutzen das Binärsystem:

...	128	64	32	16	8	4	2	1
	2^7	2^6	2^5	2^4	2^3	2^2	2^1	2^0

Beispiel

$$\begin{array}{cccccccccccc} 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 \\ 1 \times 2^{10} & 1 \times 2^9 & 1 \times 2^8 & 1 \times 2^7 & 1 \times 2^6 & 0 \times 2^5 & 1 \times 2^4 & 1 \times 2^3 & 1 \times 2^2 & 0 \times 2^1 & 1 \times 2^0 \end{array}$$

Von Dezimal- zu Binärzahlen

Umwandeln von Dezimalzahl 333 in Binärzahl:

Von Dezimal- zu Binärzahlen

Rechenbeispiel: Umwandeln von Dezimalzahl 333 in Binärzahl:

Dezimal	333 : 2 =	166	Rest 1	
	166 : 2 =	83	Rest 0	
	83 : 2 =	41	Rest 1	
	41 : 2 =	20	Rest 1	
	20 : 2 =	10	Rest 0	
	10 : 2 =	5	Rest 0	
	5 : 2 =	2	Rest 1	
	2 : 2 =	1	Rest 0	

Dual ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓

1 0 1 0 0 1 1 0 1

Von Dezimal- zu Binärzahlen

Umwandlung der Binärzahl 1101101_2 in eine Dezimalzahl:

Von Dezimal- zu Binärzahlen

Umwandlung der Binärzahl 1101101, in eine Dezimalzahl:

1	1	0	1	1	0	1
2^6	2^5	2^4	2^3	2^2	2^1	2^0
$1 * 2^6$	$1 * 2^5$	$0 * 2^4$	$1 * 2^3$	$1 * 2^2$	$0 * 2^1$	$1 * 2^0$
64	32	0	8	4	0	1

$$64 + 32 + 8 + 4 + 1 = 109$$

$$2^0 = 1$$

Kleines 1x1 der Bits und Bytes

1 Byte	= 8 Bits			
1 KB	= 1 Kilobyte	= 10^3 Bytes	=	1'000 Bytes
1 MB	= 1 Megabyte	= 10^6 Bytes	=	1'000'000 Bytes
1 GB	= 1 Gigabyte	= 10^9 Bytes	=	1'000'000'000 Bytes
1 TB	= 1 Terabyte	= 10^{12} Bytes	=	1'000'000'000'000 Bytes

Referenzen

- Hansen, H.R., Neumann, G. (2001): Wirtschaftsinformatik I
- Mit Inhalten von Matthias Baumgartner, Katharina Reinecke und Tom Schreiber

Credits

- ?
- Icon Super Mono by Double-J
- Image On-Off-Switch by All Vectors
- Image Hörfläche und Farbmodelle: wikipedia
- Image Soundwave-vector by vectorstock
- Hansen, H.R., Neumann, G. (2001): Wirtschaftsinformatik I
- Hidden:
 - Image Smoke signals by sundve (www.flickr.com/photos/sundve/3766277273)
 - Font AlphaSmoke by Beeline

Weiss nicht, welche
davon noch aktuell sind

Referent

Prof. Abraham Bernstein
Institut für Informatik

© Universität Zürich
Digital Society Initiative



Sie möchten dieses Video weiterverwenden?
Hier geht es zu den Nutzungsbedingungen...

6.2 Digitalisierung von Text



Universität
Zürich UZH

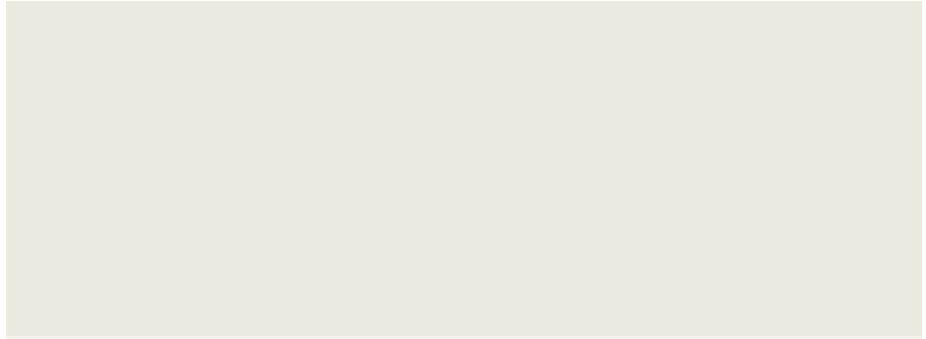
Digital Society Initiative



Studium Digitale

Kursbaustein 2: Digitalisierung von Information

Lektion 2: Digitalisierung von Text



Prof. Abraham Bernstein

UZH, Institut für Informatik und Digital Society Initiative



Diese Lektion:

- Einleitung
- **Digitalisierung von Text**
- Digitalisierung von Ton
- Digitalisierung von Bild

Digitalisierung von Text: Nicht alle Verfahren sind geeignet

A



Icon adapted; By Jack Cai, CC BY 3.0, softicons.com

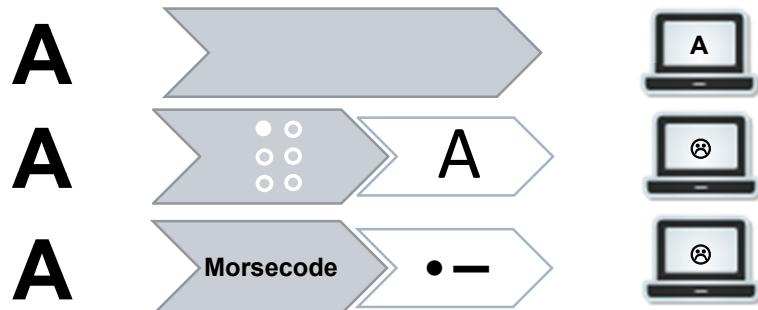
Digitalisierung von Text: Nicht alle Verfahren sind geeignet

A

A



Icon adapted; By Jack Cai, CC BY 3.0, softicons.com

Digitalisierung von Text: Nicht alle Verfahren sind geeignet

Icon adapted; By Jack Cai, CC BY 3.0, softicons.com

Digitalisierung von Text: ASCII-Code

American Standard Code for Information Interchange

ASCII-Codetabelle für Zahlen und Zeichen

0-9: 48-57 A-Z: 65-90 a-z: 97-122

7 Bit: Reicht für 128 Zeichen (2^7)

→ OK für alle Zeichen der englischen Sprache

Icon adapted; By Jack Cai, CC BY 3.0, softicons.com

Digitalisierung von Text: ASCII-Code



American Standard Code for Information Interchange

ASCII-Codetabelle für Zahlen und Zeichen

0-9: 48-57 A-Z: 65-90 a-z: 97-122

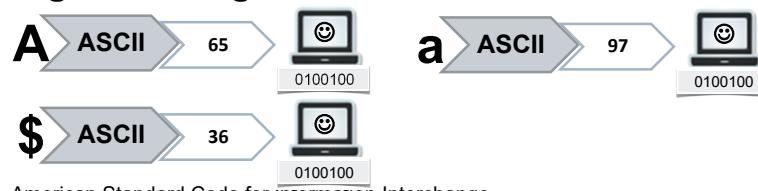
7 Bit: Reicht für 128 Zeichen (2^7)

→ OK für alle Zeichen der englischen Sprache

→ Ungenügend u.a. für europäische Sprachen

Icon adapted; By Jack Cai, CC BY 3.0, softicons.com

Digitalisierung von Text: ASCII-Code



American Standard Code for Information Interchange

ASCII-Codetabelle für Zahlen und Zeichen

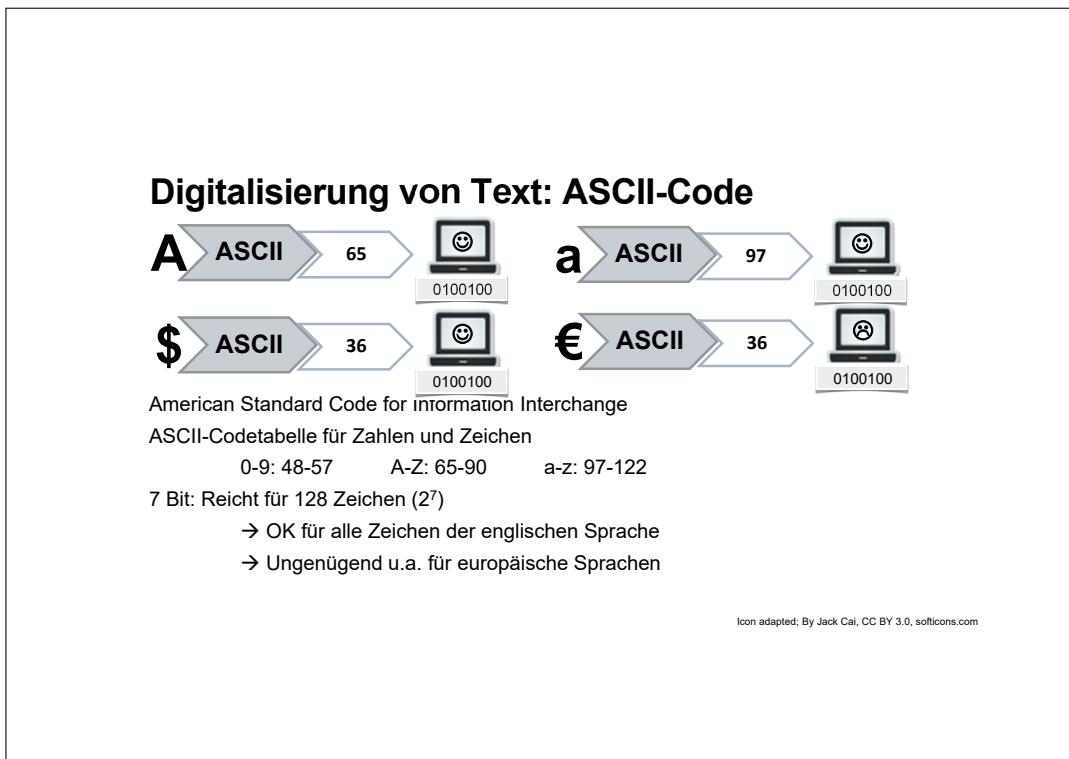
0-9: 48-57 A-Z: 65-90 a-z: 97-122

7 Bit: Reicht für 128 Zeichen (2^7)

→ OK für alle Zeichen der englischen Sprache

→ Ungenügend u.a. für europäische Sprachen

Icon adapted; By Jack Cai, CC BY 3.0, softicons.com



Digitalisierung von Text: ASCII-Code Tabelle

Decimal	Hexadecimal	Binary	Octal	Char	Decimal	Hexadecimal	Binary	Octal	Char	Decimal	Hexadecimal	Binary	Octal	Char
0	0	00000000	0		48	30	11000000	60	0	96	60	11000000	140	
1	1	00000001	1		49	31	11000001	61	1	97	61	11000001	141	
2	2	00000010	2		50	32	11000010	62	2	98	62	11000010	142	b
3	3	00000011	3		51	33	11000011	63	3	99	63	11000011	143	c
4	4	00000100	4		52	34	11000100	64	4	100	64	11000100	144	d
5	5	00000101	5		53	35	11000101	65	5	101	65	11000101	145	e
6	6	00000110	6		54	36	11000110	66	6	102	66	11000110	146	f
7	7	00000111	7		55	37	11000111	67	7	103	67	11000111	147	g
8	8	00001000	10		56	38	11000100	70	8	104	68	11000100	150	h
9	9	00001001	11		57	39	11000101	71	9	105	69	11000101	151	i
10	A	00001010	12		58	3A	11000110	72	:	106	6A	11000110	152	j
11	B	00001011	13		59	3B	11000111	73	;	107	6B	11000111	153	k
12	C	00001100	14		60	3C	11000110	74	<	108	6C	11000110	154	l
13	D	00001101	15		61	3D	11000111	75	=	109	6D	11000111	155	m
14	E	00001110	16		62	3E	11000100	76	>	110	6E	11000100	156	n
15	F	00001111	17		63	3F	11000101	77	?	111	6F	11000101	157	o
16		00000000	20		64	40	10000000	100	@	112	70	11000000	160	p
17		00000001	21		65	41	10000001	101		113	71	11000001	161	q
18		00000010	22		66	42	10000010	102	B	114	72	11000010	162	r
19		00000011	23		67	43	10000011	103	C	115	73	11000011	163	s
20		00000100	24		68	44	10000100	104	D	116	74	11000100	164	t
21		00000101	25		69	45	10000101	105	E	117	75	11000101	165	u
22		00000110	26		70	46	10000110	106	F	118	76	11000110	166	v
23		00000111	27		71	47	10000111	107	G	119	77	11000111	167	w
24		00001000	28		72	48	10000100	108	H	120	78	11000100	168	x
25		00001001	29		73	49	10000101	109	I	121	79	11000101	169	y
26		00001010	30		74	4A	10000110	110	J	122	7A	11000110	170	z
27		00001011	31		75	4B	10000111	111	K	123	7B	11000111	171	{
28		00001100	32		76	4C	10001000	114	L	124	7C	11001000	174	
29		00001101	33		77	4D	10001001	115	M	125	7D	11001001	175]
30		00001110	34		78	4E	10001110	116	N	126	7E	11001110	176	_
31		00001111	35		79	4F	10001111	117	O	127	7F	11001111	177	{DEU}
32		00000000	36		80	50	10000000	118						
33		00000001	37		81	51	10000001	121	Q					
34		00000010	38		82	52	10000010	122	R					
35		00000011	39	#	83	53	10000011	123	S					
36		00000100	40	\$	84	54	10000100	124	T					
37		00000101	41	%	85	55	10000101	125	U					
38		00000110	42	*	86	56	10000110	126	V					
39		00000111	43	&	87	57	10000111	127	W					
40		00001000	44	^	88	58	10000000	130	X					
41		00001001	45	_	89	59	10000001	131	Y					
42		00001010	46	-	90	5A	10000010	132	Z					
43		00001011	47	+	91	5B	10000011	133	{					
44		00001100	48	*	92	5C	10000100	134]					
45		00001101	49	#	93	5D	10000101	135	_					
46		00001110	50	\$	94	5E	10000000	136	~					
47		00001111	51	^	95	5F	10000001	137	-					

Source: Wikimedia Commons <https://commons.wikimedia.org/wiki/File:ASCII-Table.svg>

Digitalisierung von Text: ISO 8859-15



8 Bit: ISO 8859 definiert 256 Zeichen (2⁸)

Die ersten 128 Zeichen stimmen mit ASCII überein

Nicht geeignet für arabische, chinesische, griechische, hebräische, japanische, etc.

Schriften

Icon adapted; By Jack Cai, CC BY 3.0, softicons.com

Digitalisierung von Text: Unicode



32 Bit: Über 4 Mrd. Zeichen (2³²) können codiert werden

Grundsätzlich können alle Alphabete der Erde codiert werden

Z.Z. werden über 100'000 Zeichen erfasst (Unicode 6.2)

Auf Klingonisch muss weiterhin verzichtet werden: Unicode-Konsortium verweigert die Aufnahme hartnäckig

Icon adapted; By Jack Cai, CC BY 3.0, softicons.com

Digitalisierung von Text: Unicode

Es existieren verschiedene weitere Codierungssysteme – passend für die jeweiligen Bedürfnisse

Je universeller das System, desto mehr Bits werden für die Codierung benötigt → Speicherbedarf!

Damit die Zeichen korrekt interpretiert werden können, ist die Angabe des verwendeten Codierungssystems zwingend

Icon adapted; By Jack Cai, CC BY 3.0, softicons.com

Je mehr Text, desto höher der Speicherbedarf

1 Seite Text

80 Zeichen/Zeile

64 Zeilen/Seite

8 Bit pro Zeichen (= 1 Byte, ISO 8859-15)

$$= 80 \times 64 \times 8 = 40'960 \text{ Bits}$$

$$\rightarrow 40'960 / 8 = 5'120 \text{ Bytes} \simeq 5,1 \text{ kB}$$

Speicherbedarf: 5,1 kB

Axelsson, Bernhard und Nathalie. *Not Is This Really Science? The Semantic Web's Guide to Evaluating Research Contributions*. 2014 Technical Report.

standing on a platform and the train stops, then we will be left with three figures in the platform. Note that we have now removed words from the top (including the first one) and added three new ones. However, the last is just a pure textual construct and we have decided whether it is correct. In the rest of this paper, we will focus on the first two cases, where the question is whether it is correct and, more important, of extracting sentences and others that are truly such.

Now what about stamps? Stamp-collecting is a great tradition in science. Even though some interpret Birkhoff's statement as something derogatory towards stamp-collecting, it is still a good example of how we can gather observations about "The World" and we usually organize our structures around them. For example, we can collect all the stamps that people communicate about things. Just consider the International Classification of Diseases (ICD) or the International Standard Book Number (ISBN) which organizes books and the Standard Industrial Classification (SIC) Code which organizes companies. These are just examples of how we can collect things in the world, but how do we know that we selected our collection "correctly"? How do we demonstrate that our organization is not only "correct", but also helps others? This is the main idea behind the Semantic Web. We can collect information in projects that, at least on the surface, appear to be stamp collecting. There cannot be any guarantee that the collected information is correct. This applies not only for the goal of collecting them (i.e., we are observing the world not of texts), but also, like there are many potential sources contradicting one another.

With being other "stamps" or "stamp collectors", our overarching goal is obviously to collect as much information as possible. This is why, in general, we want others to be able to use our theories and maybe to build upon them. The idea is that we can collect as much information as possible and then publishing and making it available for reuse. Many have relied on this theory since Stigler's work on the history of regression analysis.

In both approaches, there is a need to establish a notion of correctness of our results. In the first case, we can just compare our results with a few lab experiments, we typically gather evidence supporting our cause and often accept it as correct. In the second case, we can just compare our results with what we believed following this procedure were correct. In other words, the results must be consistent with our expectations. This is what we mean by pure consistency. These are questions of reliability and validity. Let's do a few examples.

<http://www.ncbi.nlm.nih.gov/contributorinfo/registration/index.html>
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC166160/> or <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC166160/>

Referenzen

Icons: Jack Cai, Double-J Design, CC-BY 3.0

Bilder:

- <https://commons.wikimedia.org/wiki/File:ASCII-Table.svg>
- <https://commons.wikimedia.org/wiki/File:Hoerflaeche.svg>
- Mehrere basierend auf <https://commons.wikimedia.org/wiki/File:Digital.signal.svg>
- <https://commons.wikimedia.org/wiki/File:AdditiveColorMixingII.png>

Inhalte

- Hansen, H.R., Neumann, G. (2001): Wirtschaftsinformatik I

Das sind die von
mir nun erfassten

Credits

- ?
- Icon Super Mono by Double-J
- Image On-Off-Switch by All Vectors
- Image Hörfläche und Farbmodelle: wikipedia
- Image Soundwave-vector by vectorstock
- Hansen, H.R., Neumann, G. (2001): Wirtschaftsinformatik I
- Hidden:
 - Image Smoke signals by sundve (www.flickr.com/photos/sundve/3766277273)
 - Font AlphaSmoke by Beeline

Weiss nicht, welche
davon noch aktuell sind

Referenzen

- Hansen, H.R., Neumann, G. (2001): Wirtschaftsinformatik I
- Mit Inhalten von Matthias Baumgartner, Katharina Reinecke und Tom Schreiber

Referentin

Prof. Abraham Bernstein
Institut für Informatik

© Universität Zürich
Digital Society Initiative



Sie möchten dieses Video weiterverwenden?
Hier geht es zu den Nutzungsbedingungen...

6.3 Digitalisierung von Ton



Universität
Zürich UZH

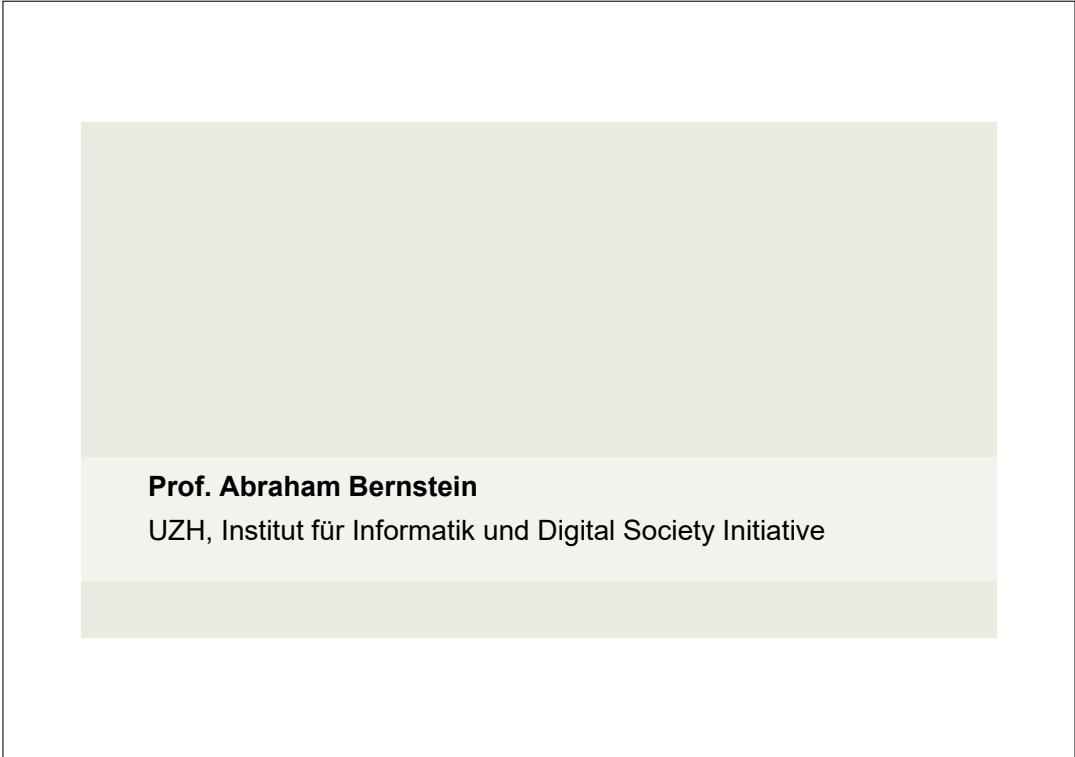
Digital Society Initiative



Studium Digitale

Kursbaustein 2: Digitalisierung von Information

Lektion 3: Digitalisierung von Ton



Prof. Abraham Bernstein

UZH, Institut für Informatik und Digital Society Initiative



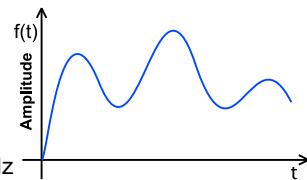
Diese Lektion:

- Einleitung
- Digitalisierung von Text
- **Digitalisierung von
Ton**
- Digitalisierung von Bild

Exkurs: Schalleigenschaften

Frequenz (Tonhöhe)

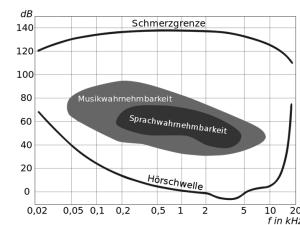
- 1 Hz = eine Schwingung pro Sekunde
- Je höher die Frequenz, desto höher der Ton
- Der Mensch hört zwischen 20 Hz und 22'000 Hz



<https://commons.wikimedia.org/w/index.php?title=Hörfläche.svg&oldid=16208133>
basiert auf <https://commons.wikimedia.org/w/index.php?title=Digital.signal.svg>

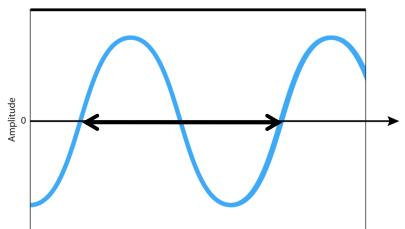
Lautstärke

- 60 dB: Unterhaltung
- 90 dB: Laute Musik
- 120 dB: Flugzeug



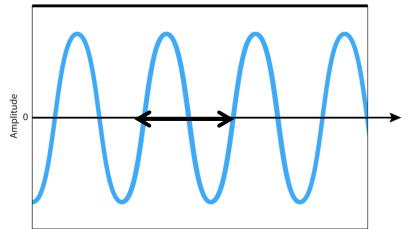
Exkurs: Schalleigenschaften

- Frequenz (Tonhöhe)
 - 1 Hz = eine Schwingung pro Sekunde

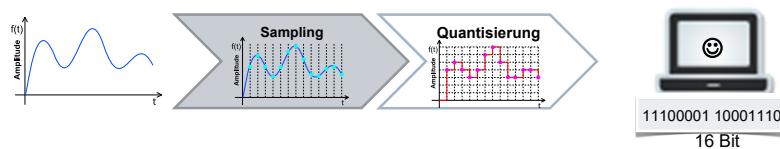


Exkurs: Schalleigenschaften

- Frequenz (Tonhöhe)
 - 1 Hz = eine Schwingung pro Sekunde
 - Je höher die Frequenz, desto höher der Ton



Tondigitalisierung: Sampling und Quantisierung



Abtasten (Sampling)

→ Periodisches Abtasten des analogen Signals

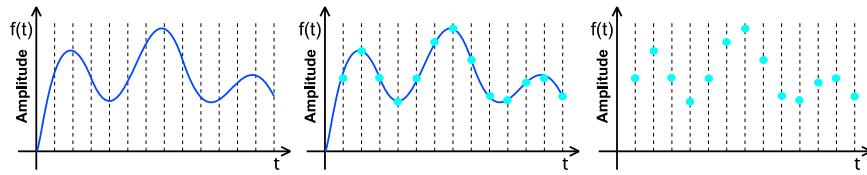
Quantisierung

→ Zuordnen der ermittelten Werte an vorgegebene Werte

Basierend auf <https://commons.wikimedia.org/wiki/File:Digital.signal.svg>
Icon adapted by Jack Cai, CC BY 3.0, sonicons.com

Sampling

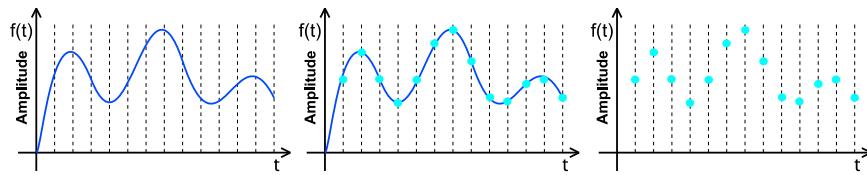
- Analoges Signal wird in regelmässigen Zeitabständen gemessen
- Gemessene Werte ergeben Abbild des Ursprungssignals



basierend auf <https://commons.wikimedia.org/wiki/File:Digital.signal.svg>

Sampling

Je höher die Zahl der Samples pro Sekunde (= Abtastrate), desto besser entspricht das digitale Abbild dem Original

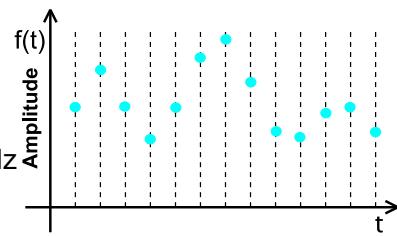


basierend auf <https://commons.wikimedia.org/wiki/File:Digital.signal.svg>

Sampling

Für gute Ergebnisse muss die Abtastrate *mindestens das Doppelte* der erwünschten maximalen Frequenz betragen

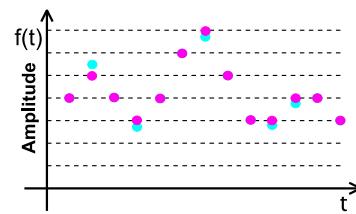
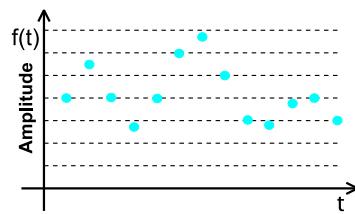
Da der Mensch maximal 22'000 Hz hört, sollte die Abtastrate mindestens 44'000 Hz betragen
(CD-Qualität: 44'100 Hz)



basiert auf <https://commons.wikimedia.org/wiki/File:Digital.signal.svg>

Quantisierung

Erfasste Messwerte werden gerundet und einem vordefinierten Wert zugewiesen

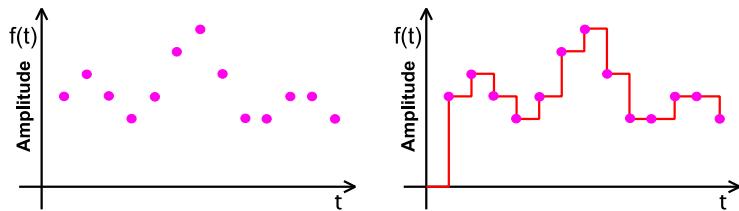


basiert auf <https://commons.wikimedia.org/wiki/File:Digital.signal.svg>

Quantisierung

Durch Hinzufügen eines Bits wird das Ergebnis doppelt so genau

- 8 Bits: 256 unterscheidbare Tonwerte
- 16 Bits: 65'536 unterscheidbare Tonwerte



basiert auf <https://commons.wikimedia.org/wiki/File:Digital.signal.svg>

**Je besser die Tonqualität,
desto höher der
Speicherbedarf**

1 Minute Musik (CD-Qualität)

- 60 Sekunden
- 44'100 Samples pro Sekunde
- 16 Bit pro Sample (= 2 Bytes)
- Stereo = 2 Kanäle
- = $60 \times 44'100 \times 2 \times 2$
- = 10'584'000 Bytes
- ➔ Speicherbedarf: 10,6 MB

Referenzen

Icons: Jack Cai, Double-J Design, CC-BY 3.0

Bilder:

- <https://commons.wikimedia.org/wiki/File:ASCII-Table.svg>
- <https://commons.wikimedia.org/wiki/File:Hoerflaeche.svg>
- Mehrere basierend auf <https://commons.wikimedia.org/wiki/File:Digital.signal.svg>
- <https://commons.wikimedia.org/wiki/File:AdditiveColorMixingII.png>

Inhalte

- Hansen, H.R., Neumann, G. (2001): Wirtschaftsinformatik I

Das sind die von
mir nun erfassten

Credits

- ?
- Icon Super Mono by Double-J
- Image On-Off-Switch by All Vectors
- Image Hörfläche und Farbmodelle: wikipedia
- Image Soundwave-vector by vectorstock
- Hansen, H.R., Neumann, G. (2001): Wirtschaftsinformatik I
- Hidden:
 - Image Smoke signals by sundve (www.flickr.com/photos/sundve/3766277273)
 - Font AlphaSmoke by Beeline

Weiss nicht, welche
davon noch aktuell sind

Referenzen

- Hansen, H.R., Neumann, G. (2001): Wirtschaftsinformatik I
- Mit Inhalten von Matthias Baumgartner, Katharina Reinecke und Tom Schreiber

Referentin

Prof. Abraham Bernstein
Institut für Informatik

© Universität Zürich
Digital Society Initiative



Sie möchten dieses Video weiterverwenden?
Hier geht es zu den Nutzungsbedingungen...

6.4 Digitalisierung von Bild



Universität
Zürich UZH

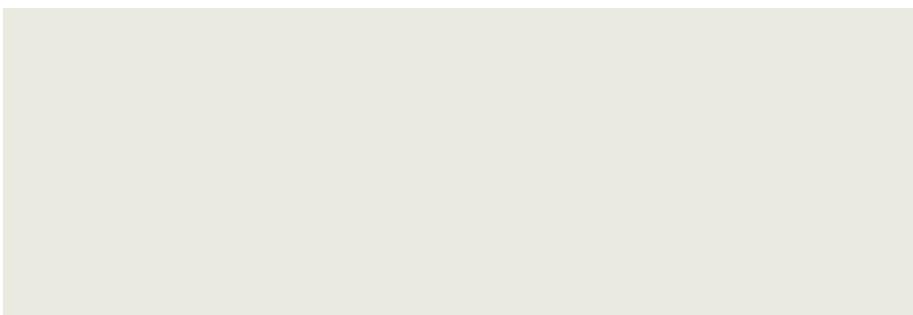
Digital Society Initiative



Studium Digitale

Kursbaustein 2: Digitalisierung von Information

Lektion 4: Digitalisierung von Bild



Prof. Abraham Bernstein

UZH, Institut für Informatik und Digital Society Initiative



Diese Lektion:

- Einleitung
- Digitalisierung von Text
- Digitalisierung von Ton
- **Digitalisierung von Bild**

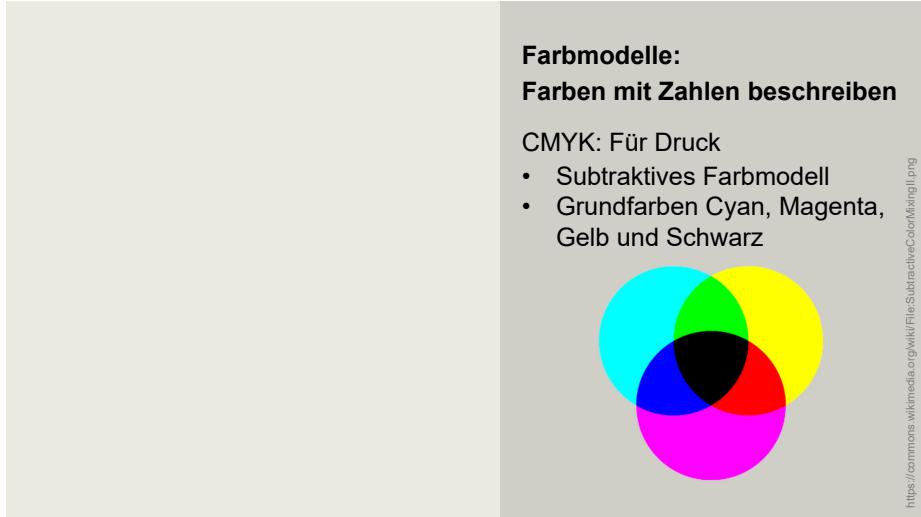


Farbmodelle:
Farben mit Zahlen beschreiben

RGB: Für Bildschirme

- Additives Farbmodell
- Drei Grundfarben:
Rot, Grün, Blau

[https://commons.wikimedia.org/wiki/File/AdditiveColorMixing.png](https://commons.wikimedia.org/wiki/File>AdditiveColorMixing.png)



The diagram shows a large light gray rectangle on the left and a smaller dark gray rectangle on the right. Inside the dark gray rectangle, there is a circular diagram of four overlapping colored circles: cyan, magenta, yellow, and black. The overlapping areas create various colors, demonstrating the subtractive color mixing process used in printing.

Farbmodelle:
Farben mit Zahlen beschreiben

CMYK: Für Druck

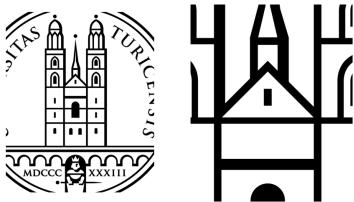
- Subtraktives Farbmodell
- Grundfarben Cyan, Magenta, Gelb und Schwarz

<https://commons.wikimedia.org/wiki/File/SubtractiveColorMixing.png>



Vektorgrafik

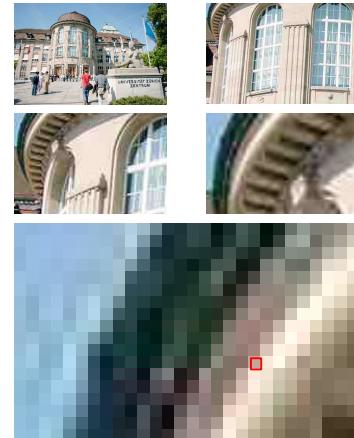
- Besteht aus mathematisch definierten Objekten
- Jedes Objekt hat Attribute. Es werden nur Attributwerte gespeichert
- Jedes Objekt kann einzeln bearbeitet werden
- Vergrössern ohne Qualitätseinbusse



The diagram illustrates the difference between vector and raster graphics. On the left is the official seal of the University of Zurich, which is a complex vector graphic. On the right is a simplified, blocky representation of the same seal, which is a raster graphic. This comparison demonstrates that vector graphics can be scaled infinitely without losing quality, unlike raster graphics which become pixelated when enlarged.

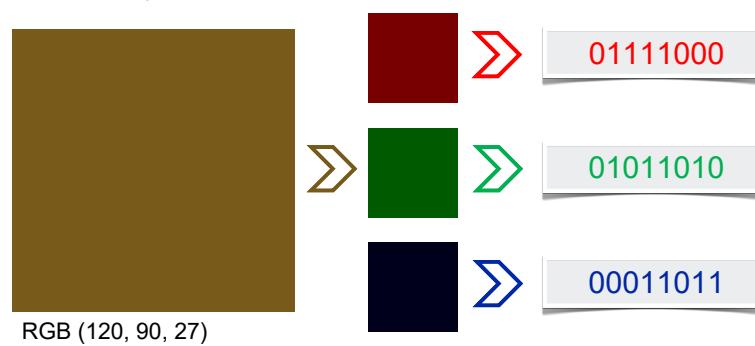
Pixelgrafik

- Besteht aus rasterförmig angeordneten Punkten (Pixel)
- Jeder Punkt hat Farbwert gemäss Farbmodell
- Jeder Punkt kann einzeln bearbeitet werden



Pixelgrafik: Farben

Beispiel RGB, 24 Bit Farbtiefe: Für jedes Pixel 8 Bit pro Grundfarbe → 256 Farben pro Kanal → 16.7 Mio. Farben



Pixelgrafik: Speicherbedarf

Beispiel RGB, 24 Bit Farbtiefe: Für jedes Pixel 8 Bit pro Grundfarbe →
256 Farben pro Kanal → 16.7 Mio. Farben



RGB (120, 90, 27)



Pixelgrafik: Speicherbedarf

Postkarte



1748×1240 Pixel = 2'167'520 Pixel

×

011110000101101000011011

=

24 Bit pro Pixel

52'020'480 Bit = 6'502'560 Bytes

Speicherbedarf: 6,5 MB

Icon adapted: By Jack Cai, CC BY 3.0, softicons.com

Fazit

- Information kann in verschiedenen Formen vorkommen (z.B. Text, Bild, Ton)
- Für die informatikgestützte Verarbeitung müssen die Informationen aufbereitet und in Binärdaten umgewandelt werden
- Detailtreue Erfassung vergrössert den Speicherbedarf

Referenzen

Icons: Jack Cai, Double-J Design, CC-BY 3.0

Bilder:

- <https://commons.wikimedia.org/wiki/File:ASCII-Table.svg>
- <https://commons.wikimedia.org/wiki/File:Hoerflaeche.svg>
- Mehrere basierend auf <https://commons.wikimedia.org/wiki/File:Digital.signal.svg>
- <https://commons.wikimedia.org/wiki/File:AdditiveColorMixingII.png>

Inhalte

- Hansen, H.R., Neumann, G. (2001): Wirtschaftsinformatik I



Das sind die von mir nun erfassten

Credits

- ?
- Icon Super Mono by Double-J
- Image On-Off-Switch by All Vectors
- Image Hörfläche und Farbmodelle: wikipedia
- Image Soundwave-vector by vectorstock
- Hansen, H.R., Neumann, G. (2001): Wirtschaftsinformatik I
- Hidden:
 - Image Smoke signals by sundve (www.flickr.com/photos/sundve/3766277273)
 - Font AlphaSmoke by Beeline



Weiss nicht, welche davon noch aktuell sind

Referenzen

- Hansen, H.R., Neumann, G. (2001): Wirtschaftsinformatik I
- Mit Inhalten von Matthias Baumgartner, Katharina Reinecke und Tom Schreiber

Referentin

Prof. Abraham Bernstein
Institut für Informatik

© Universität Zürich
Digital Society Initiative



Sie möchten dieses Video weiterverwenden?
Hier geht es zu den Nutzungsbedingungen...

6.5 **Datenspeicherung**



Universität
Zürich UZH

Digital Society Initiative



Studium Digitale

Kursbaustein 2: Digitalisierung von Information

Lektion 5: Datenspeicherung

Prof. Abraham Bernstein
UZH, Institut für Informatik und Digital Society Initiative

Diese Lektion:

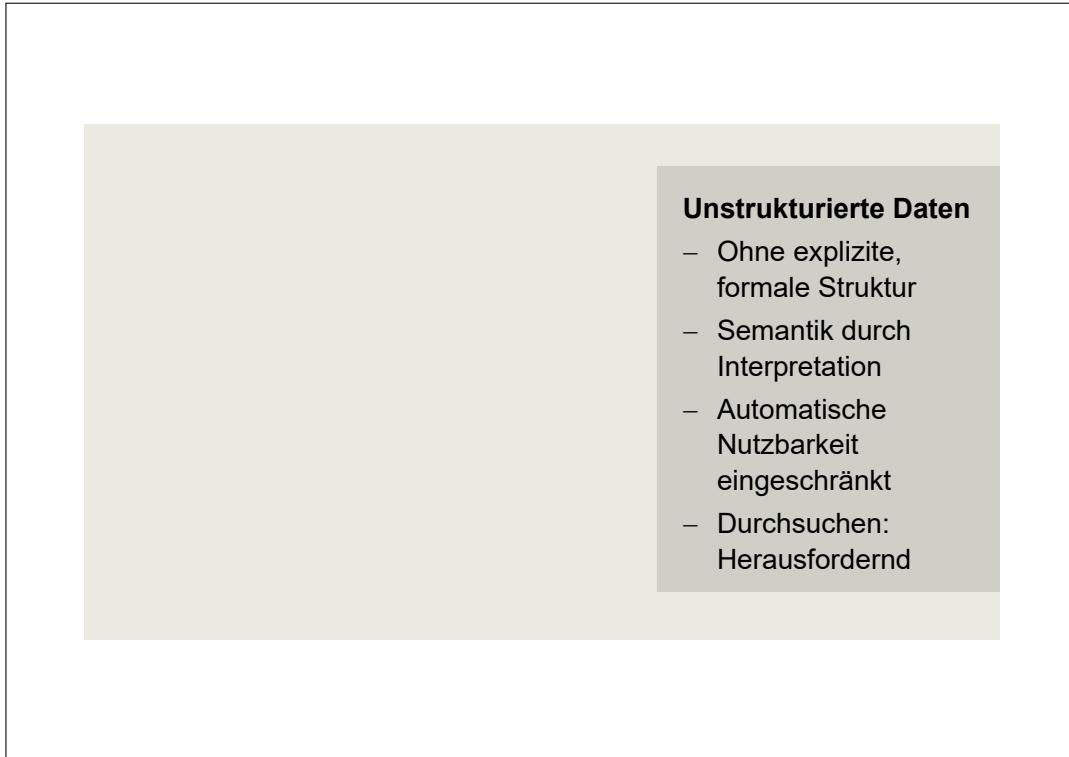
Datenspeicherung
– Unstrukturierte Daten
– Strukturierte Daten

Diese Lektion:

- Datenspeicherung
- **Unstrukturierte Daten**
 - Strukturierte Daten

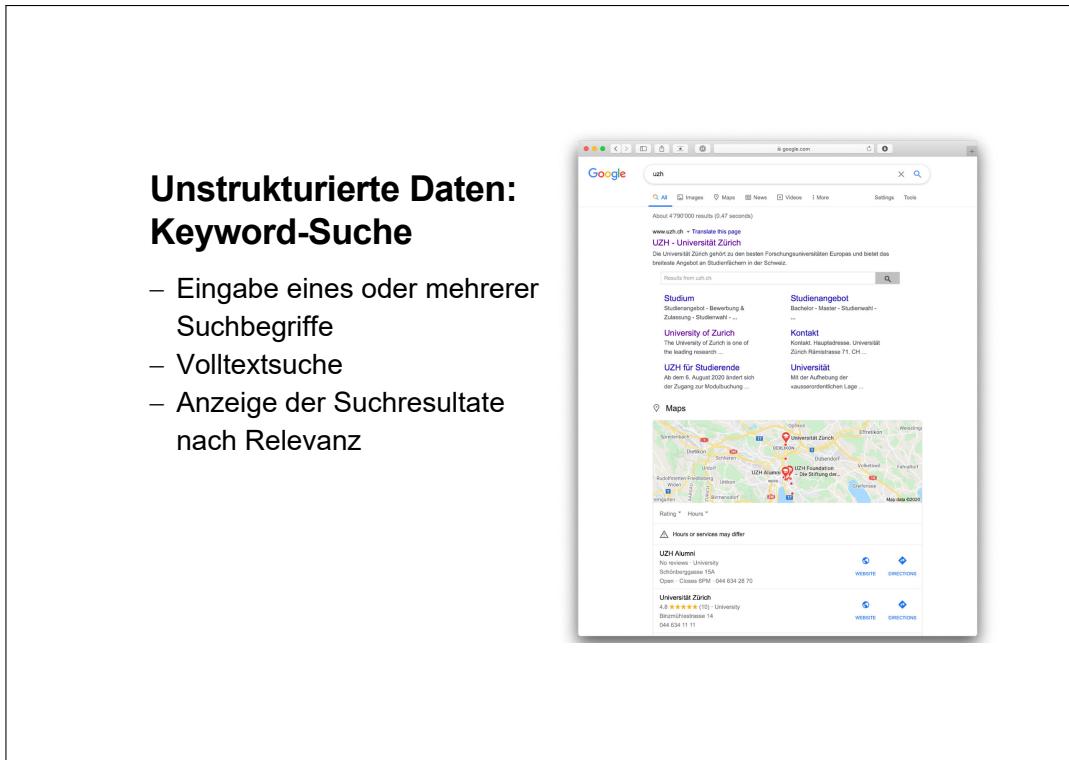
Getting information off the Internet is like taking a drink from a fire hydrant.

Mitch Kapor



Unstrukturierte Daten

- Ohne explizite, formale Struktur
- Semantik durch Interpretation
- Automatische Nutzbarkeit eingeschränkt
- Durchsuchen: Herausfordernd



Unstrukturierte Daten: Keyword-Suche

- Eingabe eines oder mehrerer Suchbegriffe
- Volltextsuche
- Anzeige der Suchresultate nach Relevanz

The screenshot shows a Google search results page for the query "uzh". The top result is a link to the University of Zurich's website. Below the link, there is a snippet of text from the website. To the right of the snippet, there is a map showing the location of the University of Zurich in Zurich, Switzerland. At the bottom of the snippet, there are two cards: one for "UZH Aarau" and one for "University Zürich".

Results from uzh.ch

Study
Studienangebot - Bewerbung & Zulassung - Studienwahl -

University of Zurich
The University of Zurich is one of the leading research ...

UZH für Studierende
Ab dem 6. August 2020 ändern sich der Zugang zur Mediothek ...

Contact
Kontakt - Hauptstrasse, Universität Zürich Rämistrasse 71, CH -

University
Mit der Aufhebung der causerenrechtlichen Lage ...

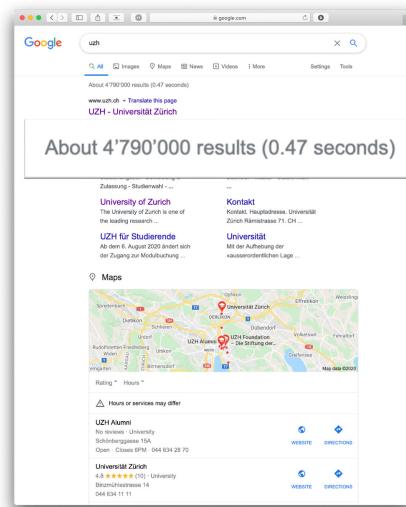
Maps

UZH Aarau
No license - University
Schlierenstrasse 14
Office phone: +41 44 634 28 70

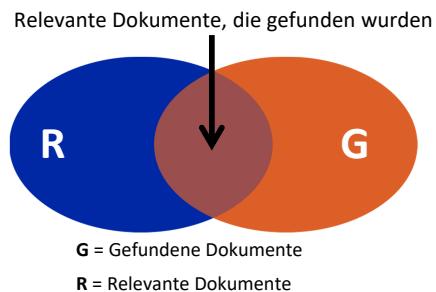
University Zürich
4.8 ★★★★★ (12) - University
Blumenstrasse 14
844 634 11 11

Unstrukturierte Daten: Keyword-Suche

- Eingabe eines oder mehrerer Suchbegriffe
- Volltextsuche
- Anzeige der Suchresultate nach Relevanz



Probleme der Keyword-Suche

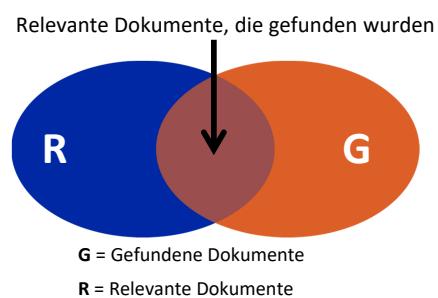


Probleme der Keyword-Suche

- Keyword-Suche findet viele irrelevante Dokumente

$$\text{Precision} = \frac{|R \cap G|}{|G|}$$

P = 100%
Alle gefundenen Dokumente sind relevant



Probleme der Keyword-Suche

- Keyword-Suche findet viele irrelevante Dokumente

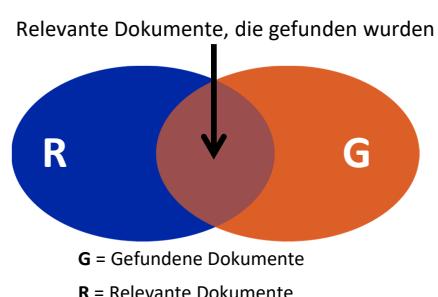
$$\text{Precision} = \frac{|R \cap G|}{|G|}$$

P = 100%
Alle gefundenen Dokumente sind relevant

- Keyword-Suche findet nicht alle relevanten Dokumente

$$\text{Recall} = \frac{|R \cap G|}{|R|}$$

R = 100%
Alle relevanten Dokumente wurden gefunden



Wie kann man diese Problematik angehen?

- Semantik
- Personalisierung
- Page Rank

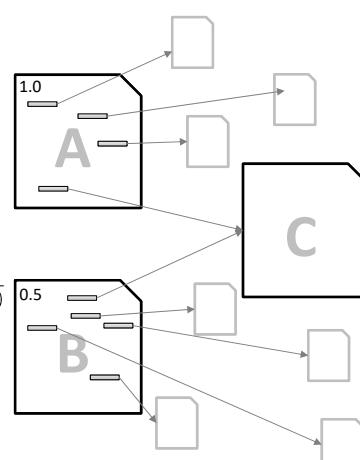
Page Rank

Versucht anhand der Linkstruktur die Wichtigkeit von Dokumenten zu erschliessen

$$\text{PageRank}_A = (1 - d)$$

$$+ d * \sum_{n \in N} \frac{\text{PageRank}_n}{\text{Anzahl(AusgehendeLinks}_n)}$$

d = Dämpfungsfaktor (z.B. 0.85)
 N = Dokumente mit Link nach A



Page Rank von C berechnen

$$d = 0.85$$

$$N = \{A, B\}$$

A: PageRank = 1.0, 4 ausgehende Links

B: PageRank = 0.5, 5 ausgehende Links

$$\text{PageRank}_C = (1 - d)$$

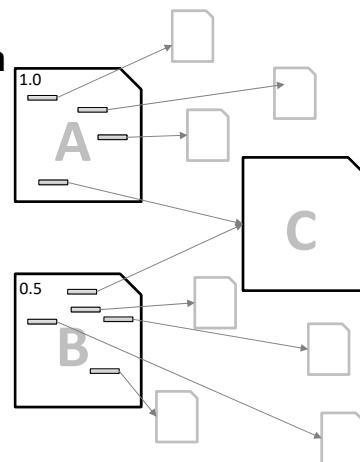
$$+ d * \sum_{n \in N} \frac{\text{PageRank}_n}{\text{Anzahl(AusgehendeLinks}_n)}$$

$$= (1 - 0.85)$$

$$+ 0.85 * (\frac{1.0}{4} + \frac{0.5}{5})$$

$$= 0.15 + 0.85 * (0.25 + 0.1)$$

$$= 0.4475$$



Diese Lektion:

Datenspeicherung

- Unstrukturierte Daten
- **Strukturierte Daten**

Strukturierte Daten

- Haben explizite, formaler Struktur
- Semantik vorgegeben
- Automatische Nutzbarkeit möglich
- Durchsuchen: Problemlos

Strukturierte Daten

Daten werden strukturiert in einer Datenbank gespeichert

- Beziehungen durch mathematische Relationen modelliert

Verwaltung, z.B., mittels Datenbankmanagementsystemen

Suche mit logikbasierter Sprache

- Relevanz gemäss Suchkriterien
- Zusätzliche Ergebnisse dank Nutzung von Relationen

Von Unstrukturierten zu Strukturierte Daten

Petra Graf, Tulpenweg 14 in Zürich, hat am 24. August die Bücher Prozessstudien zur Golfstromablösung in windgetriebenen numerischen Modellen von Joachim Dengg sowie die Geschichte des Golfstroms und seiner Erforschung von Johann Georg Kohl ausgeliehen. Andreas Zenk aus Bern (Brunnenstrasse 27) hat sich am 28. August das Buch *Golf: Zehn Lektionen in der Kunst des Scheiterns* des Autors Gregor Eisenhauer sowie die DVD *Golf für Dummies* ausgeborgt. Matthias Schwab, der am Idaplatz 7 in Bremgarten wohnt, sollte das *Golf Schrauberhandbuch - Reparieren und optimieren leicht gemacht* (von Lindsay Porter) schon vor einer Woche zurückbringen (23 Juni ausgeliehen).

Von Unstrukturierten zu Strukturierte Daten

Petra Graf wohnt am Tulpenweg 14 in Zürich. Sie hat das Buch „Prozessstudien zur Golfstromablösung in windgetriebenen numerischen Modellen“ des Autors Joachim Dengg am 24. August ausgeliehen. Sie hat das Buch „Die Geschichte des Golfstroms und seiner Erforschung“ des Autors Johann Georg Kohl am 24. August ausgeliehen.

Andreas Zenk wohnt an der Brunnenstrasse 27 in Bern. Er hat das Buch „Golf: Zehn Lektionen in der Kunst des Scheiterns“ des Autors Gregor Eisenhauer am 28. August ausgeliehen. Er hat die DVD „Golf für Dummies“, am 28. August ausgeliehen.

Matthias Schwab wohnt am Idaplatz 7 in Bremgarten. Er hat das Buch „Das Golf Schrauberhandbuch - Reparieren und optimieren leicht gemacht“ der Autorin Lindsay Porter am 23 Juni ausgeliehen.

Von Unstrukturierten zu Strukturierte Daten

Petra Graf wohnt am Tulpenweg 14 in Zürich. Sie hat das Buch „Prozessstudien zur Golfstromablösung in windgetriebenen numerischen Modellen“ des Autors Joachim Dengg am 24. August ausgeliehen. Sie hat das Buch „Die Geschichte des Golfstroms und seiner

des Autors Joh **Person** **Strasse** **Ort** **Medienart** **Titel**
Andreas Zenk wohnt an der Brunnenstrasse 27 in Bern. Er hat das Buch „Golf: Zehn Lektionen in der Kunst des Scheiterns“ des Autors Gregor Eisenhauer am 28. August ausgeliehen. Er hat die DVD „Golf für Dummies“ am **Autor** **Datum**

Matthias Schwab wohnt am Idaplatz 7 in Bremgarten. Er hat das Buch „Das Golf Schrauberhandbuch - Reparieren und optimieren leicht gemacht“ der Autorin Lindsay Porter am 23 Juni ausgeliehen.

Von Unstrukturierten zu Strukturierte Daten

Person Andreas Zenk wohnt an der **Strasse** Brunnenstrasse 27 in **Ort** Bern. Er hat das **Medienart** Buch **Titel** Golf: Zehn Lektionen in der Kunst des Scheiterns des Autors **Autor** Gregor Eisenhauer am **Datum** 28. August ausgeliehen. Er hat die **Medienart** DVD **Titel** „Golf für Dummies“ am **Datum** 28. August ausgeliehen.

Von Unstrukturierten zu Strukturierte Daten

Person_Start

Name_Start Andreas Zenk Name_Ende wohnt an der Strasse_Start Brunnenstrasse 27

Strasse_Ende in Ort_Start Bern Ort_Ende.

Er hat das Medienart_Start Buch Medienart_Ende Titel_Start Golf: Zehn Lektionen in der Kunst des Scheiterns Titel_Ende des Autors Autor_Start Gregor Eisenhauer Autor_Ende am Datum_Start 28. August Datum_Ende ausgeliehen.

Er hat die Medienart_Start DVD Medienart_Ende Titel_Start „Golf für Dummies, Titel_Ende am Datum_Start 28. August Datum_Ende ausgeliehen.

Person_Ende

Von Unstrukturierten zu Strukturierte Daten

```
<Person>
  <Name>Andreas Zenk </Name>
  <Strasse>Brunnenstrasse 27 </Strasse>
  <Ort>Bern </Ort>
  <Ausleihe>
    <Datum>28. August </Datum>
    <Medienart>Buch </Medienart>
    <Titel>Golf: Zehn Lektionen in der Kunst des Scheiterns </Titel>
    <Autor>Gregor Eisenhauer </Autor>
  </Ausleihe>
  <Ausleihe>
    <Datum>28. August </Datum>
    <Medienart>DVD </Medienart>
    <Titel>Golf für Dummies </Titel>
  </Ausleihe>
</Person>
```

Von Unstrukturierten zu Strukturierte Daten

```
<Person>
  <Name>Andreas Zenk </Name>
  <Strasse>Brunnenstrasse 27 </Strasse>
  <Ort> Bern </Ort>
  <Ausleihe>
    <Datum> 28. August </Datum>
    <Medienart> Buch </Medienart>
    <Titel> Golf: Zehn Lektionen in der Kunst des Scheiterns </Titel>
    <Autor> Gregor Eisenhauer </Autor>
  </Ausleihe>
  <Ausleihe>
    <Datum> 28. August </Datum>
    <Medienart> DVD </Medienart>
    <Titel> Golf für Dummies </Titel>
  </Ausleihe>
</Person>
```

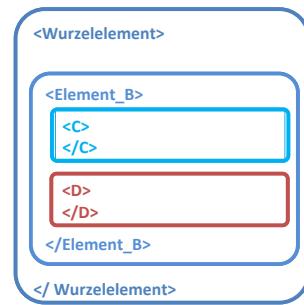
- XML (Extensible Markup Language) ist ein (menschenlesbares) Datenformat
- Häufig zum Datenaustausch verwendet
- Trägt einen Teil der Strukturinformation mit sich

Struktur eines XML Dokuments

- Ein Wurzelement
- Wurzelement kann weitere Elemente beinhalten
- Richtig verschachtelt:
Jedes Kindelement muss geschlossen sein, bevor das Elternelement geschlossen wird

Struktur eines XML Dokuments

- Ein Wurzelement
- Wurzelement kann weitere Elemente beinhalten
- Richtig verschachtelt:
Jedes Kinderelement muss geschlossen sein, bevor das Elternelement geschlossen wird



Fazit

- Viele unstrukturierte Daten → Probleme
- Struktur vereinfach Verarbeitung
- XML ist eine Strukturierungs-möglichkeit

Fazit

Viele Daten liegen unstrukturiert vor

Eine Volltextsuche liefert viele irrelevante Ergebnisse

Verfahren wie PageRank, Personalisierung etc. versuchen, die Relevanz der Suchresultate zu erhöhen

Strukturierte Daten werden basierend auf einem Datenmodell in Datenbanken gespeichert

Ein Datenbankmanagementsystem ermöglicht die effiziente Datenbankbewirtschaftung

SQL-Abfragen (Datenbankabfragen) durchsuchen auch riesige Datenbanken schnell und liefern relevante Ergebnisse

XML ermöglicht es, Daten mit Strukturinformationen zu erfassen

Referenzen

- Hansen, H.R., Neumann, G. (2001): Wirtschaftsinformatik I
- Mit Inhalten von Matthias Baumgartner, Katharina Reinecke und Tom Schreiber

Referent

Prof. Abraham Bernstein
Institut für Informatik

© Universität Zürich
Digital Society Initiative



Sie möchten dieses Video weiterverwenden?
Hier geht es zu den Nutzungsbedingungen...

6.6 Datenformate



Universität
Zürich UZH

Digital Society Initiative



Studium Digitale

Kursbaustein 2: Digitalisierung von Information

Lektion 6: Datenformate

Prof. Abraham Bernstein
UZH, Institut für Informatik und Digital Society Initiative

Diese Lektion:

- Datenformate
- Digital Rights Management
- Datenkompression

Diese Lektion:

- **Datenformate**
- Digital Rights Management (DRM)
- Datenkompression

Dateien haben
verschiedene Formate:

DOC doc, html, odt, rtf, txt,
xml, ...

MUSIC aac, dts, mp3, ogg, wav,
wma, ...

PIC Pixel: bmp, gif, jpg, png,
raw, tiff, ...

Vektor: eps, svg, ...

VIDEO: avi, H.264, mov,
mp4, ...

By Jack Cai, CC BY 3.0, softicons.com

Formate:

- Interpretation der Inhalte
- Werden über oft Endungen identifiziert
- Werden durch Hersteller oder Gremien definiert.

Formate unterscheiden sich bezüglich:

- Struktur
(Syntax und Semantik)
- Möglichkeit für Digital Rights Management
- Datenkompression

Diese Lektion:

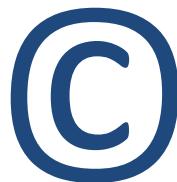
- Datenformate
- **Digital Rights Management (DRM)**
- Datenkompression



The content of this slide could not be displayed because the associated license is not valid.

Please contact your assistant to purchase a valid license key.

Digital Rights Management (DRM)



- Definition
 - Digitales Management von Rechten, um die Verbreitung und Nutzung digitaler Güter durch technische Massnahmen zu kontrollieren
- Ziel
- Durchsetzen von Urheberrechten
- Übertragen bestehender Geschäftsmodelle auf digitale Welt

Digital Rights Management (DRM)

- Definition
 - Digitales Management von Rechten, um die Verbreitung und Nutzung digitaler Güter durch technische Massnahmen zu kontrollieren
- Ziel
- Durchsetzen von Urheberrechten
- Übertragen bestehender Geschäftsmodelle auf digitale Welt

Vier Rechte stehen typischerweise im Fokus



- **Transportrechte**
Kopieren, Übertragen, Verleihen
- **Vorführrechte**
Anzeigen, Ausdrucken, Abspielen
- **Sicherungsrechte**
Schutz vor Beschädigung oder Verlust
- **Editierrechte**
Verändern, Extrahieren

Nicht mehr das digitale Gut, sondern die Nutzung kostet



- Rechteinhaber vergibt Nutzungslicenz
- Definiert, *wer, was, wann, wie, und wie oft innerhalb welcher Zeitperiode zu welchem Preis* mit den Daten machen darf:
 - Access Control
 - Usage Control
 - Content Tracking
 - Payment Management
- Lizenznehmer hat das Recht, das Gut gemäss den vereinbarten Konditionen zu nutzen

Kritik

drm is
drm is **good**
drm is **bad**
drm is **evil**
drm is **stupid**

Google

- Kein einheitlicher Standard
→ Kompatibilitätsprobleme
- Systeme nicht sicher
→ DRM kann umgangen werden
- Wettbewerb wird geschwächt
→ Marktzutrittsschranken
- Datenschutz
- Kosten vs. Nutzen

Diese Lektion:

- Datenformate
- Digital Rights Management (DRM)
- **Datenkompression**

The diagram shows a large empty box on the left and a smaller box on the right containing text and icons.

Je multimedialer, desto grösser:

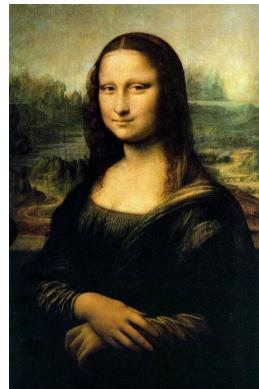
- 1 Seite Text 5.1 kB
- 1 Bild: 6.5 MB
- 1 Min Ton: 10.6 MB
- 1 Min Film: 1.3 GB

Datenkompression verringert Datengrösse
→ weniger Platz
→ schnellere Übertragung

By Jack Cai, CC BY 3.0, softicons.com

The illustration features the Mona Lisa painting on the left. A blue thought bubble originates from her hand, containing the text: "Darf ich die Mona Lisa hier verwenden?". To the right of the thought bubble, the word "Zeichen" is written vertically.

Datenkompression: verlustfreie Verfahren



- Redundante Daten werden durch kürzere Zeichen ersetzt.

– Musterwiederholung (z.B. Lauflängenkodierung)

AAAAAAABBBBAAAAAA

6A4B5A



Datenkompression: verlustfreie Verfahren



- Redundante Daten werden durch kürzere Zeichen ersetzt.

– Musterwiederholung (z.B. Lauflängenkodierung)

AAAAAAABBBBAAAAAA

6A4B5A



- Reversibel: Daten lassen sich wieder Bit-genau herstellen

Datenkompression: verlustbehaftete Verfahren



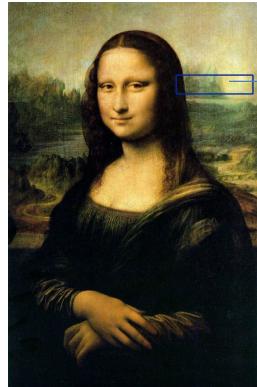
- Nutzen die *Begrenztheit menschlicher Sinne*
 - Helligkeit wird stärker registriert als Farbe oder Farbsättigung
 - Bestimmte Ton-Frequenzen werden nicht wahrgenommen

Datenkompression: verlustbehaftete Verfahren



- Nutzen die *Begrenztheit menschlicher Sinne*
 - Helligkeit wird stärker registriert als Farbe oder Farbsättigung
 - Bestimmte Ton-Frequenzen werden nicht wahrgenommen
- Kompression entfernt Teile des Originals

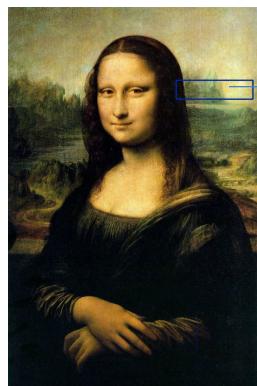
Datenkompression: verlustbehaftete Verfahren



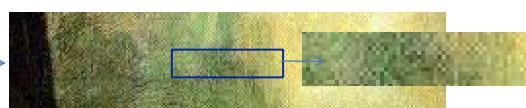
- Nutzen die *Begrenztheit menschlicher Sinne*
 - Helligkeit wird stärker registriert als Farbe oder Farbsättigung
 - Bestimmte Ton-Frequenzen werden nicht wahrgenommen
- Kompression *entfernt* Teile des Originals
- *Nicht reversibel*: Rekonstruktion entspricht nur näherungsweise den Originaldaten



Datenkompression: verlustbehaftete Verfahren



- Nutzen die *Begrenztheit menschlicher Sinne*
 - Helligkeit wird stärker registriert als Farbe oder Farbsättigung
 - Bestimmte Ton-Frequenzen werden nicht wahrgenommen
- Kompression *entfernt* Teile des Originals
- *Nicht reversibel*: Rekonstruktion entspricht nur näherungsweise den Originaldaten



Fazit

- Zahlreiche Formate
- Einige ermöglichen eine Kontrolle der Nutzung
- Dateigröße kann durch Kompression verringert werden

Referenzen

- Hansen, H.R., Neumann, G. (2001): Wirtschaftsinformatik I
- Mit Inhalten von Matthias Baumgartner, Katharina Reinecke und Tom Schreiber

Referent

Prof. Abraham Bernstein
Institut für Informatik

© Universität Zürich
Digital Society Initiative



Sie möchten dieses Video weiterverwenden?
Hier geht es zu den Nutzungsbedingungen...