

EVISU - Energievisualisierung
Literaturrecherche – Data Science

Auftraggeber
Bundesamt für Energie
Rolf Moser

Auftragnehmer
Hochschule Luzern
Technik & Architektur
Zentrum für Integrale Gebäudetechnik ZIG
Technikumstrasse 21
CH-6048 Horw

Verfasser
Reto Marek
Curdin Derungs
Stefan Winterberger
Hochschule Luzern – Technik & Architektur

Inhaltsverzeichnis

1	Einleitung	3
1.1	Begriffe im Bereich der «Data Science»	3
1.2	Strukturierung dieses Dokuments.....	5
2	Daten-Aufbereitung	7
2.1	Zeitstempel	7
2.2	Mess-Intervall.....	8
2.3	Daten-Bereinigung	9
2.4	Metadaten	11
3	Statistische Analysen.....	12
3.1	Deskriptive Statistik	12
3.2	Inferenzstatistik	13
3.3	Multivariate Analysen	13
4	Explorative Datenanalyse	15
5	Zeitreihen-Analysen	16
5.1	Zeitreihenplots.....	16
5.2	Transformationen	16
5.3	Glättung.....	17
5.4	Dekomposition	17
5.5	Auto-Korrelation	18
5.6	Kreuz/Kross-Korrelation	19
5.7	Ausreisser-Detektion	19
6	Visual Analytics	20
6.1	Visual Analytics Prozess	21
6.2	Visuelle Daten-Exploration.....	23
7	Daten-Visualisierung	25
7.1	Datengetriebener Design Prozess	25
7.2	Narrative Daten-Visualisierungen (Data Storytelling)	26
7.3	Feedback mit Ziel Verhaltensänderung.....	27
7.4	Visualisierungen des Energieverbrauchs.....	27
8	State of the Art Tools.....	28
9	Anhang	30
9.1	Literaturempfehlungen	30
9.2	Visualisierungs-Galerie	32
10	Literaturverzeichnis	58

1 Einleitung

1.1 Begriffe im Bereich der «Data Science»

Oft werden «Data Science» und «Data Analytics» fälschlicherweise austauschbar verwendet, ebenso «Data Analytics» und «Data Mining». Einleitend werden deshalb häufig verwendete Begriffe rund um «Data Science» kurz erläutert und anschliessend der Umfang dieser Literaturrecherche eingeschränkt. Wenn nicht explizit erwähnt stammen die Begriffsdefinitionen von Wikipedia oder Niebler (2018).

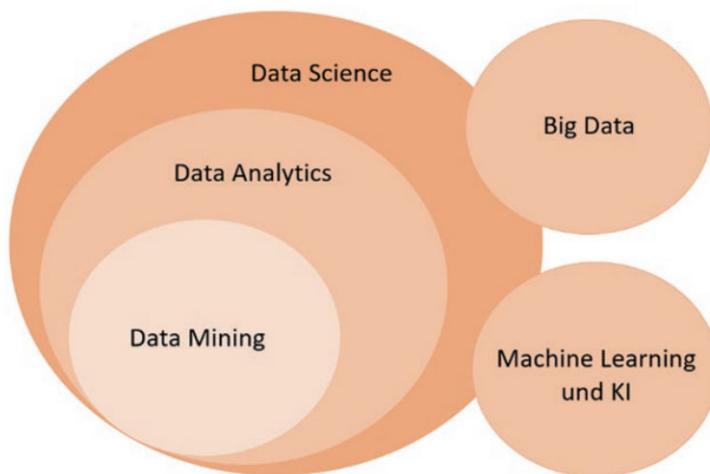


Abb. 1: Zusammenhang der Begrifflichkeiten im Bereich Datenanalyse, Quelle: Niebler (2018)

Daten

Zum Zweck der Verarbeitung zusammengefasste Zeichen, die aufgrund bekannter oder unterstellter Abmachungen Informationen darstellen.

Big Data

Grosse Menge an Daten, welche entweder zu gross, zu komplex, zu schnell lebig oder zu schwach strukturiert sind, um sie mit manuellen und herkömmlichen Methoden der Datenverarbeitung auszuwerten. Hinsichtlich der Abgrenzung zu normalen Datenmengen werden im englischen Sprachraum die drei “V“ herangezogen: Volume, Variety und Velocity. Demnach spricht man von Big Data, wenn die Daten in hoher Menge vorliegen, in schneller Geschwindigkeit neu entstehen und wenig strukturiert sind. Für sehr grosse Datenmengen spielt der von Google entwickelte MapReduce-Ansatz eine wichtige Rolle. Dabei wird eine Aufgabe in Teilaufgaben zerlegt, die dann parallel auf vielen Rechnern verteilt bearbeitet und anschliessend wieder zu einem Gesamtergebnis zusammengeführt werden. Durch Verteilung der Arbeit auf verschiedene Computer ermöglicht dies eine sogenannte horizontale Skalierung. Grössere Mengen an Daten können dabei einfach durch weiteres Hinzukaufen von gewöhnlicher Hardware bewältigt werden, die an das entsprechende Computercluster angeschlossen wird.

Data Science

Data Science bezeichnet generell die Extraktion von Wissen aus Daten. Es vereint Techniken verschiedener Fachbereiche wie Statistik, Mathematik und Informationstechnologie. Im Kern ist damit ein Data Scientist einem Statistiker sehr ähnlich, der grundlegend dasselbe Ziel verfolgt. Neben der Statistik bedient sich die Data Science aber auch anderer Disziplinen wie Programmieren, Mathematik, Machine Learning und Prognostik. Der Begriff Data Science ist somit sehr weit gefasst und umfasst viele Teildisziplinen.

« Data Science ist die Wiederentdeckung der leistungsstarken Methoden der Datenexploration und -analyse, die von Statistikern wie John W. Tukey (1962) verwendet und gelehrt werden. Diese Methoden werden für die Zukunft der Datenanalyse revolutionär sein. »

(Donoho 2015)

Data Analytics / Data Analysis

Anwendung verschiedenster Methoden und Aktivitäten auf vorhandenen Datenbeständen mit dem Ziel, neues Wissen zu generieren. Es umfasst den gesamten Prozess vom Erfassen, Ordnen und Organisieren von Rohdaten über die Analyse zur visuellen Darstellung der Ergebnisse. Die Datenanalyse kann in deskriptive Statistik, explorative Datenanalyse (EDA) und konfirmatorische Datenanalyse (CDA) unterteilt werden. Die Datenanalyse ist spezifischer und konzentrierter als die Datenwissenschaft.

Data Mining

Data Mining ist eine Methode der Datenanalyse und bezeichnet die Anwendung statistischer Methoden auf grosse Datenbestände um beispielsweise Muster in Daten zu erkennen. Ebenso wird oft versucht, bestimmte Regelmässigkeiten zu erkennen und verborgene Querverbindungen aufzudecken. Es funktioniert jedoch nicht selbstständig, sondern benötigt immer das fachliche Wissen einer Person, welche in der Lage sein muss, die Daten richtig zu interpretieren.

Time Series Analysis

Eine Zeitreihe (Time Series) ist eine Reihe von Datenpunkten, die in zeitlicher Reihenfolge indiziert ist bzw. zu jedem Datenpunktwert einen Zeitstempel hat. Die Zeitreihenanalyse ist ein Bereich der «Data Analysis» und umfasst spezielle Methoden zur Analyse von Zeitreihendaten.

Machine Learning

Machine Learning ist ein Bereich der künstlichen Intelligenz, der Programmen die Möglichkeit gibt, mithilfe von Daten zu lernen, ohne explizit programmiert zu werden. Machine Learning weist Ähnlichkeiten zu Data Mining auf und ist ein Teilbereich des Feldes der künstlichen Intelligenz. Das Ziel von Machine Learning ist es, Strukturen in Datenbeständen anzulernen. Im Gegensatz zu Data Mining steht dabei nicht die Analyse dieser Daten im Vordergrund, sondern die Anwendung des gelernten Wissens auf neue Daten.

1.2 Strukturierung dieses Dokuments

Im Fokus dieses Projektes stehen Energiedatenvisualisierungen welche Messwerte von Sensoren als Basis haben. Dies sind normalerweise Zeitreihen, weshalb der Fokus dieser Recherche auch auf diesen Bereich zielt. Die nachfolgenden Kapitel sind wie folgt aufgeteilt:

Daten-Aufbereitung

Eine zentrale Anforderung, welche nicht Bestandteil der Forschungsfrage ist, jedoch oft vom Zeitaufwand unterschätzt wird: die zu analysierenden Daten weisen oft unterschiedliche Datenformate auf und müssen von den Analyse-Anwendungen verarbeitet werden können. Daten werden aus verschiedenen Quellen extrahiert, in ein einheitliches Format transformiert, ggf. bereinigt, ggf. aggregiert und in eine zentrale Stelle geladen, wo sie dann für weitere Analysen zur Verfügung stehen.

Statistische Analysen

Die Autoren legen für das vorliegende Projekt Wert auf die Explorative Datenanalyse und Zeitreihenanalysen. Dennoch werden Grundlagen der klassischen statistischen Analysen gegeben. Auf das Vorhersagen künftiger Werte bzw. einer Modellbildung wird in diesem Kontext verzichtet.

Explorative Datenanalyse

Die Explorative Datenanalyse (EDA) wird angewendet, um bisher unbekannte Zusammenhänge von vorhandenen Daten zu entdecken und zu verstehen. Es werden nach Mustern, Abhängigkeiten und Regelmässigkeiten gesucht um anschliessend besser beurteilen zu können, welche Daten allenfalls Ausreisser oder gar fehlerhafte Daten sind. Es handelt sich hierbei um einen iterativen Prozess, welcher ein gewisses Mass an Erfahrung des Analysten voraussetzt, um vorhandene Muster zu erkennen.

Zeitreihen-Analysen

Zeitreihen-Analysen sind in Bezug auf Energiedaten ein sehr wichtiges Mittel, da Energiedaten sehr oft mit Zeitstempeln daherkommen. Energiedaten werden meist auch genau aus dem Grund erfasst, um Trends zu erkennen und sie mit anderen artgleichen Daten zu vergleichen. Hierfür ist der Zeitstempel oft der gemeinsame Nenner. Durch die Zeitstempel können sehr schnell Aussagen darüber gemacht werden, ob Daten fehlen. Sie werden oft auch auf Zeitachsen visualisiert und können so visuell schnell verglichen werden.

Visual Analytics

Die visuelle Analytik integriert neue rechnergestützte und theoriebasierte Werkzeuge mit innovativen interaktiven Techniken und visuellen Darstellungen. Sie beschreibt einen iterativen Prozess in welchem der Mensch mit Hilfe von interaktiven Techniken und visuellen Darstellungen die Daten erkunden und aus einer Kombination von Beweisen und Annahmen Schlussfolgerungen zieht. Aus Sicht der Autoren eignet sich

dieses analytische Vorgehen für Experten in der Betriebsoptimierung und deshalb wird dieser Teil besonders tief betrachtet.

Daten-Visualisierung

Neben einfachen Darstellungen wie den Histogrammen, Streudiagrammen und Zeitreihendarstellungen entwickeln Datenwissenschaftler oft neue Darstellungsformen und Dashboards, welche das Verständnis der Daten fördert und den Fokus besser hervorhebt. Im Kontext der Präsentation und Kommunikation kommt auch die Technik des «Storytellings» vermehrt zum Einsatz.

State of the Art Tools

Um Sinnvolle Datenanalyse zu betreiben sind Informatikmitteln und geeignete Tools unerlässlich. Im professionellen Umfeld wird unterschieden zwischen Skriptsprachen mit ihren diversen Programm-Bibliotheken, um die Daten zu Analysieren und darzustellen und Plattformen, bei denen die Daten mit professionellen Dashboards untersucht werden können. Skriptsprachen sind hier weitaus flexibler, benötigen aber eine gewisse Kenntnis der entsprechenden Sprachen und Bibliotheken. Sobald Machine Learning zum Einsatz kommen soll, sind Skriptsprachen unerlässlich. Dashboards führen meistens schneller zu ersten Erkenntnissen, sind aber weniger flexibel und häufig auch kostenpflichtig. Für die End-User werden oftmals ebenfalls Dashboards für die Visualisierung der Daten eingesetzt, mit diesen kann dann aber viel weniger bis gar nicht interagiert werden. Es liegt im Ermessen der Experten geeignete Dashboards für End-User bereit zu stellen.

2 Daten-Aufbereitung

Eine der grössten Herausforderungen für Analysten wie auch Endanwender ist es, die Flut von Informationen und Daten zu erfassen und nützliche Erkenntnisse daraus zu gewinnen – kurz das Problem der Informationsüberflutung. Diese Überflutung hat neben der Zahl an verfügbaren Quellen und Datensätzen eine weitere Herausforderung, die unterschiedlichen Datenformate und Zeitstempel. Diese müssen vor einer Analyse vereinheitlicht werden. Die Datenerfassung und entsprechende Speicherung sind deshalb heute immer noch sehr zeitaufwändige Aktivitäten (Niebler et al 2018, Nielsen 2019).

Meist werden die Daten nach dem ETL-Prozess (Extract, Transform und Load) verarbeitet. Bekannt ist der Prozess aus den 1970-er Jahren vor allem durch die Verwendung in «Daten-Warenhäusern». Grosse Datenmengen werden aus mehreren unterschiedlichen Quellen *extrahiert* und in ein einheitliches Format *transformiert*, um dann ins «Data-Warehouse» *geladen* zu werden.

Ein richtig konzipiertes ETL-System ... (Wikipedia 2020)

- extrahiert Daten aus verschiedenen Quellsystemen,
- setzt Datenqualitäts- und Konsistenz Standards durch,
- passt Daten so an, dass getrennte Quellen gemeinsam genutzt werden können und
- liefert resp. speichert Daten schliesslich in einem präsentationsreifen Format, so dass Endanwendungen und Endbenutzer Entscheidungen treffen können.

Das ETL Prinzip trifft auch auf die moderne Datenanalyse zu, denn auch hier müssen Daten aus verschiedenen Quellen zusammengeführt werden.

Nachfolgend werden wichtige Punkte aus dem ETL-Prozess beschrieben, welche aus Sicht der Autoren für das vorliegende Projekt relevant sind.

2.1 Zeitstempel

Zeitstempel sind im Rahmen von Zeitreihen zwar sehr hilfreich, können aber sehr knifflig sein und zu Problemen führen. In diesem Unterkapitel werden deshalb Schwierigkeiten von Daten mit Zeitstempeln angesprochen.

Wer hat den Zeitstempel wo erzeugt?

Nach Nielsen (2019) ist da erstens die Frage nach der Herkunft des Zeitstempels bzw. welcher Prozess oder welches Gerät diesen erzeugt hat. Nicht zwingend ist der Zeitpunkt vom Ereignis gleich dem Zeitpunkt der Aufzeichnung. Beispielsweise übermitteln IOT-Geräte teils nur Messwerte. Erst der empfangende Server im Netzwerk fügt den Zeitstempel hinzu. Dies ermöglicht den Betrieb von Geräten, ohne dass diese über eine interne Zeit verfügen. Ebenso übermitteln solche Geräte teils den Mittelwert der letzten Stunde da technologiebedingt die Anzahl an übermittelnden Daten klein gehalten werden muss. So beinhaltet der Zeitstempel von 13:10:15 beispielsweise den Mittelwert von sechs Zwischenmessungen im Zeitbereich 12:10:00 – 13:10:00. Eine gute Dokumentation der Herkunft ist nach Nielsen (2019) wünschenswert.

Zeitzone und Sommerzeit

Die meisten Zeitstempel werden in der koordinierten Weltzeit (UTC) abgespeichert welche keine Unterscheidung von Sommer- und Winter kennt. In der Informationstechnologie ist es eher unüblich, Daten mit der lokalen Zeit zu speichern. Jedoch wird in der Praxis des Gebäudemonitoring aus Erfahrung der Autoren dieser Fall immer wieder angetroffen. Ebenso kommt die Frage auf, ob die Zeitstempel die Sommer- und Winterzeitumstellung mitberücksichtigen oder nicht.

Format

Nicht standardisiert ist ebenfalls das Format wie die Zeitstempel abgespeichert werden. In einer csv-Datei werden Zeitstempel in Form von Zeichen- und Zahlenfolgen abgespeichert. Eine Software muss nun wissen, welche Zahl den Monat und welche den Tag etc. beinhaltet. Die ISO8601 (internationaler Standard der ISO, der Empfehlungen über numerische Datumsformate und Zeitangaben enthält) empfiehlt folgende Formate:

- | | |
|----------------------------|-----------------------------------|
| - 2020-08-31T10:37:55+0000 | UTC-Zeitzone |
| - 2020-08-31T10:37:55Z | UTC-Zeitzone (Z steht für UTC) |
| - 2020-08-31T12:37:55+0200 | Europe/Zurich Zeitzone Sommerzeit |
| - 2020-08-31T12:37:55+0100 | Europe/Zurich Zeitzone Winterzeit |

2.2 Mess-Intervall

Die Zeitstempel einer Zeitreihe können äquidistant, also in konstanten Intervallen (beispielsweise alle 15 Minuten, täglich etc.) oder unregelmässig angeordnet sein. Messdaten von Sensoren in der Gebäudetechnik sind oft zeitlich unausgerichtet und haben unregelmässige Intervalle. Dies weil einige Sensoren nur Wertänderungen aufzeichnen oder die Daten aus mehreren Quellen mit unterschiedlichen Einstellungen und Zeitstempeln stammen.

Für viele Daten-Analysen (wie beispielsweise in Kapitel 3) ist es notwendig, dass die zu analysierenden Daten den gleichen regelmässigen Zeit-Intervall aufweisen (Blázquez-García et al. 2020). Dies trifft auch zu, wenn man zwei unterschiedliche Zeitreihen miteinander vergleichen will.

Wenn man das Intervall beispielsweise von 15min auf 1h erhöht, spricht man von Downsampling. Man spricht hierbei auch von Aggregieren. Je nach Sensor und Anwendung werden die Daten nach unterschiedlichen Funktionen aggregiert (Minimum, Maximum, Mittelwert, Median, Summe etc.). Dem gegenüber steht das Upsampling.

Ein Downsampling kann auch aus ressourcentechnischer Sicht sinnvoll sein um Speicherplatz, Abfragezeiten und/oder Datenübertragungen zu optimieren. Wenn ein Rohdaten-Sensorwert einen 15min-Intervall hat und die Analyse bzw. Visualisierung nur eine Tagessumme verwendet, dann macht es Sinn diese Daten der bereits in der Aggregationsstufe der Anwendung ab zu speichern. Diese Reduktion kann zu verschiedenen Zeitpunkten stattfinden. Entweder beim erstmaligen Extrahieren der Daten oder in einem späteren Schritt. In einer Data Warehouse-Architektur können beispielsweise in einem ersten ETL-Schritt die Daten importiert, bereinigt und als Rohdaten abgespeichert werden. In

einem zweiten Schritt dann können die Daten aggregiert abgespeichert werden. Besonders für Zeitreihen gibt es etablierte Datenbanken wie InfluxDB, welche für das Persistieren von Zeitreihen gewöhnliche SQL-Datenbanken übertreffen. Diese Datenbanken speichern die Sensorwerte als Rohdaten. Bei der Abfrage der Datenbank kann dann die Aggregationsstufe und die Auswerte-Funktion angegeben werden (z.B. Mittelwert pro Tag). Die Datenbank gibt dann als Antwort direkt die gewünschte Aggregation zurück.

2.3 Daten-Bereinigung

Tätigkeiten der Bereinigung reichen von einer Neuformatierung der Werte bis zu aufwändiger Vorverarbeitungen wie das Entfernen/Ersetzen von Anomalien und Artefakten, Gruppierung, Glättung und Teilmengenbildung.

Der korrekte und nachvollziehbare Umgang mit nicht kompletten, inkonsistenten und fehlerhaften Daten ist äusserst wichtig. Hier gibt es unterschiedliche Handhabungen seitens Daten-Analysen, Visual Analytics und Daten-Management.

2.3.1 Fehlende Werte

Aus analytischer Sicht sollte ein fehlender Wert meistens ersetzt werden, beispielsweise durch z.B. Interpolation. Denn Ausreisser können statistische Analysen verfälschen. Dieses Vorgehen kann aber aus Datamanagement-Perspektive wichtige Fakten verschleiern; so kann anschliessend z.B. ein defekter Sensor in der Analyse nicht mehr erkannt werden. Dieses Beispiel veranschaulicht, dass je nachdem Rohwerte wie auch bereinigte Werte benötigt werden und auch so entsprechend in den Datenbanken vorhanden sein sollten.

Nielsen (2019) schlägt folgende Methoden vor, wie man mit fehlenden Daten umgehen soll:
Imputation - Füllen der fehlenden Daten anhand von Eigenschaften der Zeitreihe, z.B. durch Verwendung des letzten vorhandenen Wertes, einem gleitenden Mittelwert oder durch Interpolation der benachbarten Werten.

Interpolation - Anhand von benachbarten Datenpunkten den resp. die fehlenden berechnen.
Löschen – Betroffene Perioden oder Zeitreihen löschen und gar nicht gebrauchen. Dies ist gewiss nicht in allen Fällen ein gangbarerer Weg.

2.3.2 Inkorrekte bzw. inkonsistente Werte, Ausreisser

Hawkins (1980) definierte Ausreisser wie folgt:

«Ein Ausreisser ist eine Beobachtung, die so sehr von anderen Beobachtungen abweicht, dass der Verdacht entsteht, sie sei durch einen anderen Mechanismus hervorgerufen worden.

Bis heute gibt es gemäss Blázquez-García et al. (2020) jedoch noch keinen Konsens über die verwendeten Begriffe. Ausreisser-Beobachtungen werden oft als Anomalien, disharmonische Beobachtungen, Ausnahmen, Aberrationen, Abweichungen, Überraschungen, Eigenheiten oder Verunreinigungen bezeichnet. Der Begriff Ausreisser wird vor allem bei der Erkennung von Punktausreissern verwendet, während Anomalie häufiger bei der Erkennung von Ausreissern in Teilsequenzen verwendet wird. Man wählt den Begriff teilweise auch anhand des Ziels der Erkennung. So wird Ausreisser meist bei der Erkennung unerwünschter Daten verwendet, während Anomalie bei der Erkennung von Ereignissen verwendet wird. Da die Erkennung von Ereignissen nicht Thema der vorliegenden Arbeit ist, verwenden wir fortan nur den Begriff Ausreisser. Die Ausreisser-Erkennung gehört nach Blázquez-García et al. (2020) heute zu den wichtigsten Aufgaben des Data Minings von Zeitreihen. Im Kapitel 5.7 wird deshalb auf diese Thematik eingegangen.

Es ist jedoch Vorsicht geboten, denn die Ausreisser-Erkennung wie auch der Umgang mit fehlenden Werten haben einen Einfluss auf die statistischen Auswertungen und die Visualisierungen. Datenzentrierte Methoden zur Bereinigung können gemäss Aggarwal (2015) manchmal gefährlich sein, da sie dazu führen können, dass nützliches Wissen entfernt wird. Wichtig ist darum nach Blázquez-García et al. (2020) zu erkennen um was für Ausreisser es sich handelt:

- i) **Ungewollte Daten** - Erfordern eine Daten-Bereinigung, um die nutzlosen oder unerwünschten Daten zu entfernen. Damit wird die Datenqualität für weitere Analysen verbessert.
- ii) **«Events of interest»** - Erfordern eine Analyse, um den Grund zu erörtern.

Aggarwal (2015) unterscheidet drei Fälle:

Erkennung von Inkonsistenz - Falls Daten aus zwei verschiedenen Quellen vorhanden sind, so können inkorrekte Werte mit einem Vergleich dieser erkannt werden. Ebenso kann ggf. eine Zeitreihe mittels einem Modell eine andere Zeitreihe simulieren. Der Vergleich der simulierten und der echten Reihe kann auch inkorrekte Werte hervorbringen.

Domänen-Wissen - Ein beträchtliches Mass an Domänen-Wissen ist oft in Form der Wertebereiche, Regeln oder Modellen verfügbar. Beispielsweise ist für einen Standort der Bereich der möglichen Aussentemperaturen bekannt und somit können Temperaturen ausserhalb dieses Bereiches als unwahrscheinlich eingestuft und somit als inkorrekt taxiert werden.

Daten-Zentrierte Methoden - In diesen Fällen wird versucht Ausreisser über statistisches Verhalten der Daten zu erkennen. Näheres hierzu in Kapitel 5.7.

Abschliessend kann gesagt werden, dass die Daten-Bereinigung als Arbeitsschritt im Gesamtprozess der Daten-Analyse einen hohen Stellenwert einnimmt und noch heute ressourcen- und zeitintensiv ist.

2.4 Metadaten

Die Daten werden oft in flachen Strukturen mit unterschiedlichen Namenskonventionen abgelegt, d.h. die Daten sind über mehrere Datenquellen gesehen heterogen und nicht transparent. Gemäss Card et al. (1997) ist die sogenannte «Semantische Integration, Tagging» ein vielversprechender Ansatz, um diesem Umstand entgegen zu wirken. Hierbei werden die Datenpunkte mit definierten Tags/Metadaten versehen und einer domänenspezifischen konzeptionellen Sicht/Ontologie zugeordnet. Die Daten werden erst aussagekräftig, verständlich und nutzbar, wenn sie in den Kontext der realen Anlage gesetzt werden. Das Ziel ist es, dass die Daten selbstbeschreibend sind, ohne grossen manuellen Aufwand zur Wertschöpfung genutzt werden können und Prozesse dadurch automatisiert werden können.

Eine Community namens Project Haystack (2020) definiert in mehreren Arbeitsgruppen semantische Beschreibungen für Daten im Zusammenhang mit intelligenten Geräten und Gebäuden. Alle von der Community entwickelten Arbeiten werden als Open-Source zur Verfügung gestellt. Eine weitere Open-Source-Entwicklung in diese Richtung ist Brick Schema (2020). Das Projekt hat ebenfalls eine einheitliche Gliederung und Beschreibung der gebäudetechnischen Daten zum Ziel.

3 Statistische Analysen

Visuelle Exploration (wie sie später im Kapitel 6.2 beschrieben wird) hilft den Analysten Daten visuell zu erforschen und zu verstehen. Dies ist möglich dank der menschlichen Wahrnehmung und der Tatsache, dass der Mensch sehr gut darin ist, Muster zu erkennen, interessante und unerwartete Lösungen zu finden, Wissen aus verschiedenen Quellen zu kombinieren und im Allgemeinen kreativ zu sein (Wegner 1997). Überschreitet das zu lösende Problem aber eine gewisse Grösse, sind gemäss Aigner (2011) computergestützte Algorithmen bei numerischen Berechnungen, logischem Denken und Suchen besser (schneller und genauer). Diese halb- oder vollautomatisierten Algorithmen werden hier als «Daten-Analysen» zusammengefasst und bestehen meist aus statistischen Analysen. Diese Analysemethoden sind nicht Fachbereich spezifisch, sondern beschreiben generelle Methoden welche bislang unbekannte Zusammenhänge, Querverbindungen und Trends aus den Daten zutage führen sollen.

Nachfolgend werden verschiedene statistische Methoden erläutert, welche aus Sicht der Autoren für das Projekt von Interesse sein können.

3.1 Deskriptive Statistik

Mittels der Deskriptiven Statistik werden mit statistischen Mitteln die Daten beschrieben. Grundsätzlich wird zwischen **univarianten** (Analyse einer Variablen) und **bivarianten** (Analyse des Zusammenhangs zweier Variablen) unterschieden. Bivariate Analysen sind die Vorstufe zu multivarianten Analysen (Analyse dreier und mehr Variablen).

3.1.1 Univariate Methoden

Bei univarianten Analysen werden Häufigkeitstabellen, Diagramme sowie Mittelwert- und Streumasse erstellt und analysiert.

Beispielsweise:

- **Arithmetisches Mittel** - auch bekannt als der Durchschnitt oder Mittelwert.
Nachteil: wird durch extreme Werte und Ausreisser stark beeinflusst.
- **Median** - der zentralste Wert des Datensatzes, Teilung in zwei Hälften. 50% der Daten liegen demnach über dem Medianwert und 50% darunter.
Vorteil: wird nicht stark durch Extremwerte beeinflusst.
- **Quartile** - Teilt den Datensatz in vier Stücke. Beispiel mit dem untersten Quartil: 25% der Daten liegen unter dem Wert von $Q_{0,25}$, 75% darüber. $Q_{0,50}$ entspricht dem Median.
- **Quantile** – Werden nicht viertel verwendet, sondern eine andere Einteilung spricht man von Quantilen. Zum Beispiel 1/3 Quantile. Quartile sind also eine Spezialisierung davon.
- **Perzentile** - Analog Quartile, teilt Datensatz aber in hundert Stücke. Das 50-Perzentil ist der Median.
- **Modus/Modalwert** - Wert, der im Datensatz am häufigsten vorkommt.
- **Spannweite/Variationsbreite** - wie weit die gegebenen Daten verteilt sind, also Abstand zwischen Minimum und Maximum des Datensatzes.
Nachteil: wird durch extreme Werte und Ausreisser stark beeinflusst.

- **Quartil-Abstand** - Abstand zwischen dem 1. und 3. Quartil.
Vorteil: gegenüber Ausreisern unempfindlich.
- **Varianz/Standardabweichung** - Streuungsparameter, wie weit die Daten vom Mittelwert abweichen.
- **Normalverteilung** - Es handelt sich um einen Graphen einer Normalverteilung einer Variablen, sie wird wegen ihrer Form Glockenkurve genannt, zeigt die Verteilung um den Mittelwert.
- **Schiefe** - Sie ist das Mass für die Asymmetrie der Verteilung einer Variablen um ihren Mittelwert
- **Korrelationskoeffizienten** - Mass für den Zusammenhang zwischen zwei oder mehreren Datensätzen. Werte zwischen -1 und 1, wobei 0 keinen Zusammenhang darstellt. Werte über 0 sagen aus, dass ein Zusammenhang zwischen Datensätzen besteht (hohe Werte aus Datensatz A gehen einher mit hohen Werten von B).

3.1.2 Bivariate Methoden

Bivariate Analysen untersuchen den Zusammenhang zweier Variablen. Häufig wird zwischen abhängiger- und unabhängiger Variabel unterschieden¹. Dabei wird ausschliesslich die Stärke des Zusammenhangs, nicht aber deren Kausalität ermittelt (dies geschieht erst mittels Multivariater Analysen, siehe Kapitel 3.3).

3.2 Inferenzstatistik

Während mit der deskriptiven Statistik einzig eine Stichprobe bzw. eine Zeitreihe beschrieben wird, werden im Rahmen der schliessenden Statistik (Inferenzstatistik) mit den Resultaten dieser Stichprobe bestimmte Aussagen zur Grundgesamtheit gemacht. In der schliessenden Statistik werden Signifikanztests, Verteilungs- oder Mittelwertvergleiche vorgenommen. Mit statistischen Signifikanztest kann gesagt werden, mit welcher Wahrscheinlichkeit die Resultate einer Stichprobe auf die Grundgesamtheit übertragen werden können.

Diese Methoden werden nicht näher betrachtet, da diese aus Sicht der Autoren nicht Projektrelevant sind.

3.3 Multivariate Analysen

Stellt man mittels einer bivarianten Analyse gemäss Kapitel 3.1.2 fest, dass zwischen Zeitreihe X und Zeitreihe Y eine statistische Beziehung besteht, so kommt die Frage auf, ob da tatsächlich eine Kausalbeziehung vorhanden ist oder der Zusammenhang von einer gemeinsamen Ursache Z stammt. In multivariaten Analysen wird deshalb versucht, Drittvariablen durch statistische Techniken unter Kontrolle zu bringen. Hierzu ist erforderlich, dass eine Hypothese über den potenziellen Effekt einer Drittvariable vorliegt und Messdaten von dieser vorhanden sind.

¹ Interessiert der Einfluss der Außentemperatur auf die Innentemperatur, so ist die Außentemperatur die abhängige Variabel und die Innentemperatur die unabhängige.

Grundsätzlich werden multivariate Analysen in folgenden drei Kontexten eingesetzt, um zu beweisen, dass entweder:

- ein Zusammenhang zwischen Variablen nicht zufällig ist
(reeller Zusammenhang),
- ein indirekter Zusammenhang zweier Variablen durch eine Dritte beeinflusst ist,
sowie
- eine dritte Variable die Beziehung zwischen zwei Variablen moderiert.

Für die multivariate Analyse stehen verschiedene Verfahren wie Faktoren-, Cluster-, Hauptkomponenten- oder Regressions-Analysen zur Verfügung.

4 Explorative Datenanalyse

Die explorative Datenanalyse (EDA) beschreibt eine Erkundung der Daten (meist mit visuellen Darstellungen), von denen nur ein geringes Wissen über deren Zusammenhänge vorhanden ist. Tukey (1977) vergleicht die EDA mit Detektivarbeit: Die Arbeit eines guten Ermittlers zeichne sich dadurch aus, dass dieser wisst, wonach es sich an einem Tatort zu suchen lohnt und welche Hilfsmittel er dazu benötigt. Das Ziel der EDA ist es somit, Muster oder Auffälligkeiten zu erkennen, die neue Schlussfolgerungen ermöglichen bzw. Unbekanntes klären.

Tukey (1977) schlägt in seinem für die explorative Datenanalyse bedeutenden Standardwerk mehrere Techniken vor, welche im Folgenden aufgelistet werden.

Deskriptive Kennwerte

- Häufigkeiten (ob diese sinnvoll sind, hängt von der Art der Zeitreihe ab)
- Fünf-Zahlen-Zusammenfassung (Übersicht eines Datensatzes mit dem Minimum, unteres Quartil, Median, oberes Quartil und Maximum)
- Lagemasse (wo liegt die Verteilung, z.B. Mittelwert, Median, Quantile)
- Streuungsmassen (Varianz, Standardabweichung)

Exploration von Verteilungen

- Stamm-Blatt-Diagramme ²
- Häufigkeitsdiagramme
- Histogramme
- Boxplots

Exploration mit Streudiagrammen

- Streudiagramme (Zusammenhang zweier Variablen)
- Regressionsgeraden
- Loess (Lokale Regression)
- Streudiagramm-Matrix (Zusammenhänge von drei oder mehreren Variablen)

Multivariate Exploration

- Clusteranalysen
- Faktorenanalysen

² <https://de.wikipedia.org/wiki/Stamm-Blatt-Diagramm>. Dient der Visualisierung von Häufigkeitsverteilungen dient. Im Vergleich zu einem Boxplot oder einem Histogramm bleibt hier jedoch die Genauigkeit erhalten. Zudem lassen sich statistische Kennzahlen wie der Modalwert, Median und die Quantile abschätzen.

5 Zeitreihen-Analysen

Während die meisten Daten-Analyse-Verfahren versuchen mit möglichst allgemeinen Daten umgehen zu können, gibt es auch Spezialisierungen wie die Zeitreihen-Analyse. Hier spielen die temporalen Aspekte und Beziehungen eine grosse Rolle.

Ziele der Zeitreihenanalyse gemäss Schwarz (2013):

- Beschreibung von Zeitreihen (Sichtbarmachen innerer Zusammenhänge)
- Modellierung von Zeitreihen
- Prognose zukünftiger Werte oder Wahrscheinlichkeitsbändern

Eine wichtige Herausforderung besteht darin, Reihen mit einem ähnlichen Verlauf zu erkennen, auch wenn dieser etwas zeitlich versetzt ist, aber dennoch ähnliche Charakteristika aufweist.

5.1 Zeitreihenplots

Ein erster, wichtiger Schritt der Zeitreihenanalyse ist die simple Betrachtung des Zeitreihenplots. Die Analyse liefert Information über Trends, Saisoneffekte, Variabilität und Ausreisser. In der üblichen Darstellung ist auf der x-Achse die Zeit, auf der y-Achse der Messwert.

Indexierung

Sollen mehrere Zeitreihen mit unterschiedlichen Wertebereichen miteinander angezeigt und verglichen werden, so bietet sich die Indexierung an. Dabei werden die Werte der Zeitreihe prozentual abgebildet, wobei der erste Wert 100% oder dem Wert 1 entspricht. Somit kann die relative Entwicklung der Werte rasch miteinander verglichen werden.

5.2 Transformationen

Gemäss Dettling (2018) müssen Zeitreihen nicht unbedingt in Ihrer Rohdatenform miteinander verglichen werden. In vielen Fällen ist es besser und effizienter die Daten zu transformieren.

Lineare Transformation

Sie hat die Funktion $Y_t = a + b X_t$. Im einfachsten Fall bedeutet dies eine Änderung der Rohdaten von der Einheit Watt zu Kilowatt, was einer Division mit 1'000 entspricht ($a = 0$, $b = 0.001$). Weiter kann der Wechsel von ° Fahrenheit zu Grad Celsius betrachtet werden ($a = -32$, $b = 5/9$). Solche Transformationen ändern das Aussehen des Zeitreihenplots nicht und auch Berechnungen wie die Autokorrelation und Vorhersagemodelle werden gleich sein.

Monatliche Summen und Durchschnitte

Im Bereich der Energiedaten werden beispielsweise Wasserverbräuche oft nur monatlich ausgelesen. Werden diese Rohwerte in einem Diagramm dargestellt, so verfälschen die unterschiedlichen Tage pro Monat und auch die Schaltjahre das Bild unnötigerweise. Auch nachfolgende automatisierte Analysen werden somit verfälscht. In solchen Fällen bietet sich die Umrechnung auf spezifische Werte pro Tag an.

5.3 Glättung

Glätten bedeutet nach Definition von Wikipedia im mathematischen Kontext, eine Kurve in eine Kurve mit geringerer Krümmung zu überführen, die gleichzeitig möglichst wenig vom Original abweicht. Im Kontext von Zeitreihen-Analysen hilft Glättung, Muster und Trends besser zu sehen. Das Hervorheben von saisonalen Mustern wird in Kapitel 5.4 separat besprochen.

Das Glätten/Filtern von Daten wird nach Nielsen (2019) oft vor einer Datenanalyse durchgeführt. Einerseits kann es dazu dienen, auf eine sehr pragmatische Weise Ausreisser (siehe Kapitel 5.7) zu entfernen. Dies kann ganz einfach gesehen über einen gleitenden Mittelwert geschehen. Andererseits sind für Vorhersage-Algorithmen oft geglättete Zeitreihen von Vorteil.

Vereinfacht wird im Idealfall durch eine kleine Glättung das Rauschen entfernt, durch eine stärkere Glättung die saisonale/zyklische Komponente und so der Trend isoliert. Die Wahl der passenden Glättungsmethode ist relevant, da eine ungeeignete Methode ggf. mehr als eine Komponente auf einmal entfernt. Nachfolgend zwei Beispiele von Glättungsmethoden.

Gleitender Mittelwert

Die einfachste Methode ist der gleitende Mittelwert. Zu jedem Zeitpunkt der Reihe wird ein Mittelwert der umgebenden Werte bestimmt. Im Energie-Bereich wird beispielsweise oft die Aussentemperatur der letzten 48h gemittelt, um thermodynamische Verzögerungen des Wärmeflusses zu kompensieren.

Exponentielle Glättung

Mit der exponentiellen Glättung kann eine kurzfristige Prognose aus vergangenen Daten erstellt werden, wobei Daten aus der nahen Vergangenheit stärker gewichtet werden als ältere. Durch die Gewichtung der Zeitreihenwerte mit einem Glättungsfaktor werden starke Ausschläge einzelner beobachteter Werte auf der geschätzten Zeitreihe verteilt.

5.4 Dekomposition

Typische Zeitreihen entstehen aus dem Zusammenwirken regelhafter und zufälliger Ursachen. Die regelhaften Ursachen können periodisch (saisonale) variieren und/oder langfristige Trends enthalten. Zufällige Einflüsse werden oft als Rauschen bezeichnet. Die Dekomposition beschreibt die Zerlegung einer Zeitreihe in ihre Komponenten Trend, Saisonalität und Zufall. Oft ist es für weitere Daten-Analysen nötig, dass der Trend und die saisonale Komponente von den Rohdaten weggerechnet werden. Deshalb behandelt ein grosser Teil der Zeitreihen-Analysen entsprechende Vorgehen.

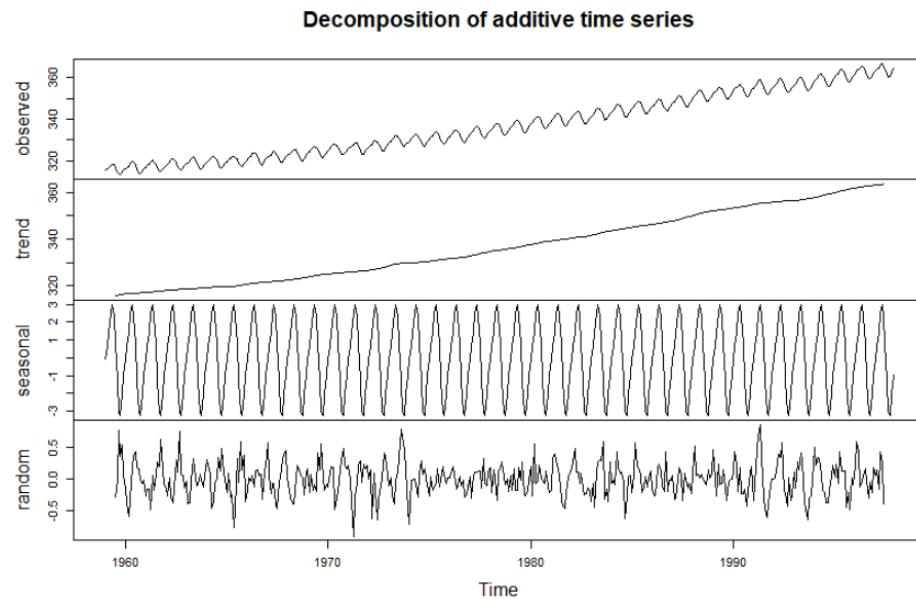


Abb. 2: Dekomposition der Mauna Loa Atmospheric CO₂ Messwerte, Quelle hslu, IGE ³

Rohdaten - In obiger Abbildung zeigt «observed» die Rohdaten der CO₂-Werte von der Messstation Mauna Loa von 1959 bis 1997. Die Werte reichen von 313ppm bis 366ppm.
Trend – Beschreibt die langfristige Entwicklung der Zeitreihe.

Saisonale Komponente – Beschreibt die saisonalen Schwankungen, in diesem Beispiel innerhalb eines Jahres. Die Werte schwanken lediglich um +/- 3ppm.

Zufall/Rest-Komponente - Einmalige und zufällige Einflüsse.

5.5 Auto-Korrelation

Die Auto-Korrelation beschreibt die Abhängigkeit von Messwerten derselben Zeitreihe zu einem früheren Zeitpunkt. D.h. die Werte einer Variable zum Zeitpunkt t werden verglichen mit Werten dieser Variable in Vorperioden t - 1, t - 2, t - 3, ...

³ Code um das Diagramm in der Statistiksoftware R selber zu erzeugen: `plot(decompose(co2))`

5.6 Kreuz/Kross-Korrelation

Das Ziel der Kreuz-Korrelation ist es, die Abhängigkeit mehrerer Zeitreihen zu beschreiben und zu verstehen. So können zeitverschobene Abhängigkeiten und Einflüsse sichtbar gemacht werden.

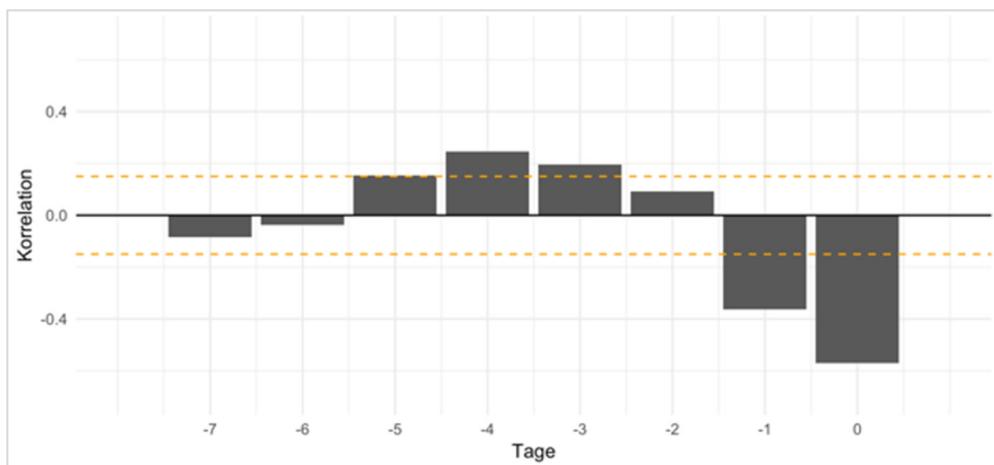


Abb. 3: Kreuz-Korrelation zwischen Heizenergieverbrauch und Außentemperatur, Quelle: hslu IGE

Als Beispiel zeigt Abb. 3 die Kreuz-Korrelation von einer gemittelten Außentemperatur und dem Heizenergieverbrauch. Damit wird ersichtlich, dass neben der aktuellen gemittelten Außentemperatur (0 auf x-Achse) auch die Temperatur des Vortages (-1) den Heizenergieverbrauch von heute beeinflusst.

5.7 Ausreisser-Detektion

Chandola *et al.* (2009) und Aggarwal (2015) stellen eine breite Übersicht über generelle Techniken der Ausreisser-Erkennung vor und Gupta *et al.* (2014) untersuchte im Speziellen Zeitreihen. Blázquez-García *et al.* (2020) zog diese Arbeit weiter und erarbeitete eine aktuelle und organisierte Übersicht über Ausreisser-Erkennungs-Techniken von Zeitreihen und listet auch Open Source Software resp. Pakete aus, mit welchen Ausreisser erkannt werden können (siehe Kapitel 0).

Nach Blázquez-García *et al.* (2020) verwenden die meisten Methoden, um Punktausreisser in Zeitreihen zu erkennen sogenannte «iterative Techniken» welche sehr rechenintensiv seien. Keine wissenschaftliche Studie habe Ausreisser-Erkennungs-Algorithmen untersucht, welche in einem «real-time» Szenario eingesetzt würden. Die erkannten Ausreisser würden in der Regel je nach Anwendung entfernt oder durch einen Erwartungswert ersetzt. Die grosse Mehrheit der Methoden schreibt anscheinend vor, dass die Zeitreihen konstante Intervalle haben (siehe Kapitel 2.2).

6 Visual Analytics

Die drei vorangehenden Kapitel beschrieben einzelne Teile eines umfangreichen Prozesses der Datenvisualisierung (Daten-Erfassung, Vorverarbeitung, Speicherung, Analyse und Visualisierung und Präsentation). «Visual Analytics» fasst diesen Prozess zusammen und beschreibt diesen als ein iteratives Vorgehen, in welchem der Mensch mit Hilfe von interaktiven Techniken und visuellen Darstellungen die Daten erkunden und aus einer Kombination von Beweisen und Annahmen Schlussfolgerungen zieht.

Gemäss Thomas und Cook (2005) ist Visual Analytics die Wissenschaft des analytischen Denkens, welches unter Zuhilfenahme von interaktiven Schnittstellen unterstützt wird. Das Menschliche Denken mit all seiner Flexibilität und der Fähigkeit Hintergrundwissen auf neue Problemstellungen anzupassen ist in diesem Gebiet unerlässlich. Visuelle Schnittstellen zu Computersystemen mit ihrer enormen Rechenleistung sollen hier den Menschen gezielt unterstützen. Das Ziel dieser Schnittstelle soll es sein, erfasste und aufbereitete Daten so darzustellen, damit sie dem Menschen einen Mehrwert bringen. So soll der Mensch auch mit der Schnittstelle interagieren können, um die für ihn relevanten Daten geeignet anzuzeigen.

Hier sind geeignete Filter unerlässlich damit der Benutzer nicht selbst komplexe Queries auf eine Datenbank machen muss. Nicht zu vernachlässigen ist vor der Visualisierung, das Erfassen der relevanten Daten aus unterschiedlichsten Quellen, diese zu filtern und zu aggregieren und geeignet zu persistieren⁴ (siehe Kapitel 2). Beim Persistieren ist darauf zu achten geeignete Strukturen zu verwenden. Besonders für Zeitreihen gibt es etablierte Datenbanken wie InfluxDB, welche für das Persistieren von Zeitreihen gewöhnliche SQL-Datenbanken übertreffen.

Das Thema Visual Analytics ist, ähnlich der Data Science, schwierig zu fassen und nur vage definiert. Es wird im Rahmen dieses Projektes darauf verzichtet, eine repräsentative Zusammenfassung des Themas wiederzugeben. Dem versierten Leser werden zwei Standardwerke zum Thema empfohlen, welche auch als Basis der nachfolgenden Abschnitte dienten (siehe Kapitel 9.1.2).

⁴ Persistenz ist in der Informatik der Begriff, der die Fähigkeit bezeichnet, Daten oder logische Verbindungen über lange Zeit bereitzuhalten (Wikipedia: Persistenz (de))

6.1 Visual Analytics Prozess

Traditionell werden gemäss Bertini et al. (2010) zwei verschiedene Datenanalyseprozesse unterschieden, welche lange separat und unabhängig erforscht wurden

- i) Automatisierte Datenanalyse
- ii) visuelle Exploration

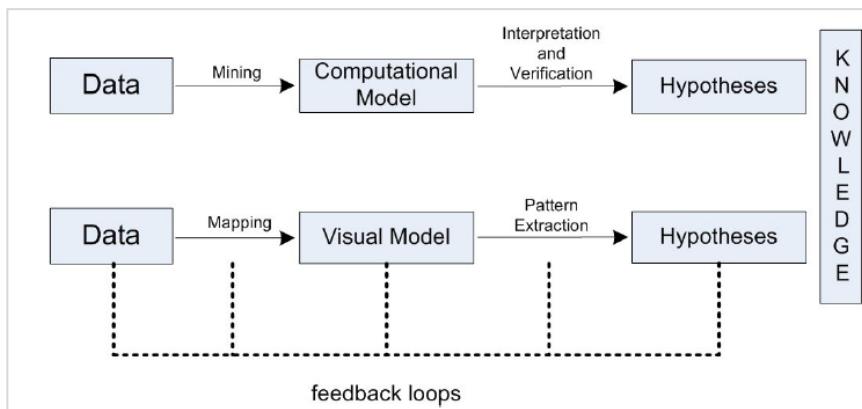


Abb. 4: Traditionelle Prozesse der Datenanalyse (Bertini et al. 2010), automatisierte Datenanalyse oben, visuelle Exploration unten

Der neu entstandene Begriff «Visual Analytics» kombiniert automatisierte und visuelle Analysetechniken. Die nachfolgende Abbildung zeigt eine abstrakte Übersicht, welche den iterativen Prozess grafisch darstellt.

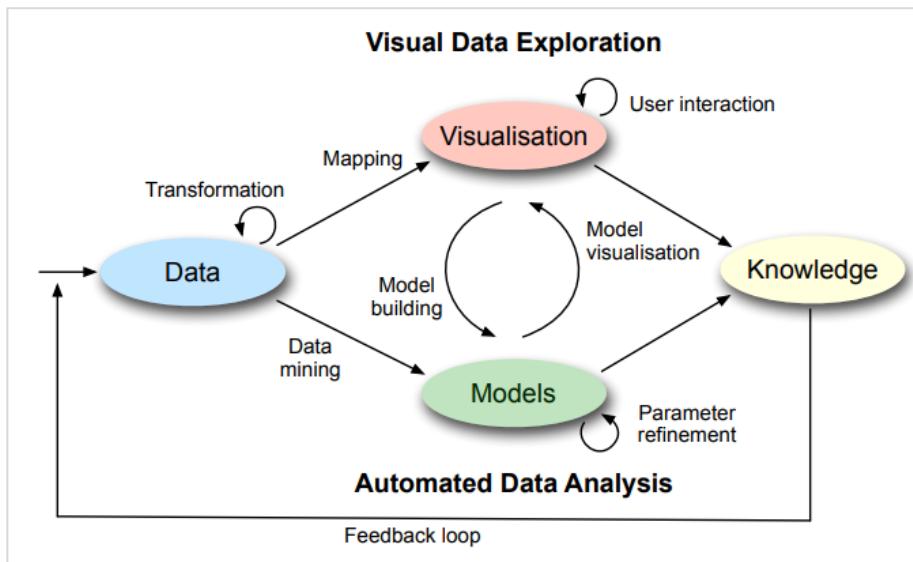


Abb. 5: abstrakte Übersicht des iterativen «Visual Analytics Prozesses» (Keim et al. 2010)

Ein generellerer Leitfaden von Schneiderman (1996) beschreibt die visuelle Erkundung von Daten wie folgt:

«Übersicht zeigen, Zoom/Filter, Details bei Bedarf»

Bei grossen Datensätzen ist es jedoch schwierig, sich als ersten Schritt eine sinnvolle Übersicht zu verschaffen. Oft gehen dadurch interessante Muster verloren. Deshalb erweitert Card et al. (1997) den Leitfaden wie folgt:

«Analysieren, Wichtiges zeigen, Zoom/Filter, weiter Analysieren, Details bei Bedarf»

Diese Anpassung verdeutlicht, dass es notwendig ist, die Daten anhand von Interesse, Fragestellungen und Hypothesen zu analysieren, die relevante Aspekte/Erkenntnisse daraus zu zeigen, um weiter zu erkunden. Dies kann oft nicht durch rein automatisierte oder rein visuelle Analysen bewerkstelligt werden und benötigt Fachwissen der zu analysierenden Domäne.

Der Prozess beginnt mit dem Verstehen der Fragestellung und der Auswahl an Daten für die Analyse. Die oft inhomogenen Daten aus verschiedensten Quellen werden in ein einheitliches Format zusammengeführt und geprüft. Dieser Schritt beinhaltet Integritätsprüfungen, Umgang mit fehlenden Daten, Normalisierung, Gruppierung etc. (siehe Kapitel 2).

Nachdem die Daten vorhanden und validiert sind, unterscheidet Keim et al. (2010) zwei Varianten für das weitere Vorgehen:

i) automatisierte Datenanalyse/Data-Mining

Hier kommen klassische Data-Mining Methoden zum Einsatz. Die anschliessende Visualisierung dient hier zur Darstellung und Auswertung der Ergebnisse, um dann weitere Analysen durch zu führen.

ii) visuelle Datenexploration

Hier versucht der Analyst Hypothesen in einem ersten Schritt visuell zu erforschen und dann anhand automatisierter Analysen zu bestätigen. Der visuelle Aspekt dient folglich der Findung geeigneter Analysemethoden und der Erkundung der Daten.

6.2 Visuelle Daten-Exploration

Im Vergleich zum Data-Mining (Kapitel 3) benötigen visuelle Datenanalysemethoden domänen spezifisches Hintergrundwissen, Kreativität und Intuition. Nachfolgend wird im Speziellen auf die Exploration und Darstellung von Zeitreihen eingegangen.

Die Datenexploration wird in der Regel mit einer Kombination aus automatisierten und manuellen Aktivitäten durchgeführt. Zu den automatisierten Aktivitäten können Daten-Profiling oder Datenvizualisierung oder tabellarische Berichte gehören, um dem Analytiker einen ersten Einblick in die Daten und ein Verständnis der Hauptmerkmale zu vermitteln (siehe Kapitel 3).

Darauf folgt oft ein manueller «Drilldown» oder Filtern der Daten, um Anomalien oder Muster zu identifizieren, die durch die automatisierten Aktionen identifiziert wurden. Die Datenexploration kann auch manuelles Skripting und Abfragen in den Daten (z.B. unter Verwendung von Sprachen wie SQL, Python oder R) oder die Verwendung von Tabellenkalkulationen oder ähnlichen Tools zur Anzeige der Rohdaten erfordern.

Andrienko und Andrienko (2006) definieren mögliche Aufgaben und Schritte eines Data Scientist, um diese zwei Fragen zu beantworten. Diese Schritte zeigen mögliche Vorgehen und Fragen, um die Daten systematisch zu erforschen.

Das Modell wird in zwei Aufgabenklassen unterteilt. **Elementare Aufgaben** beziehen sich auf einzelne Datenelemente. Dies können einzelne Werte, aber auch einzelne Datengruppen sein. Dabei geht es vor allem darum, dass die Daten separat berücksichtigt werden und nicht als Ganzes betrachtet werden. **Synoptische Aufgaben** hingegen beinhalten eine Gesamtübersicht und betrachten Wertesätze oder Datengruppen in Ihrer Gesamtheit.

Elementare Aufgaben mit Beispielen

- *Direkte Suche*
Was war die Aussentemperatur am 14. Januar um 12:00 Uhr?
- *Inverser Suche*
An welchem Tag war die Aussentemperatur am tiefsten?
- *Direkter Vergleich*
Vergleich des Heizenergieverbrauches von Haus A und B im Januar 2020.
- *Inverser Vergleich*
Hat die Aussentemperatur vor oder nach dem Sonnenaufgang 20°C erreicht?
- *Beziehungssuche*
An welchen Tagen war die Aussentemperatur höher als die Innentemperatur?

Synoptische Aufgaben

- *Direkte Suche (Musterdefinition)*
Wie war der Trend des Stromverbrauchs im Januar?
- *Inverses Nachschlagen (Mustersuche)*
In welchen Monaten ist der durchschnittliche Warmwasserverbrauch gesunken im Vergleich zum Jahresmittel?
- *Direkter (Muster-)Vergleich*
Wie ist der Warmwasserbezug im Januar verglichen mit dem Juni?
- *Inverser (Muster-)Vergleich*
Ist der Stromverbrauch abhängig von der Aussentemperatur?
- *Beziehungssuche*
Gibt es Perioden wo die Innentemperatur länger als drei zusammenhängende Tage unter 20°C war?
- *Homogenes Verhalten*
Beeinflusst die Aussentemperatur die Innentemperatur?
- *Heterogenes Verhalten*
Beeinflusst die Aussentemperatur den Stromverbrauch?

Diese theoretische Gliederung wiedergibt Fragen, welche sich ein Data-Scientist im Rahmen der visuellen Datenexploration stellen kann.

7 Daten-Visualisierung

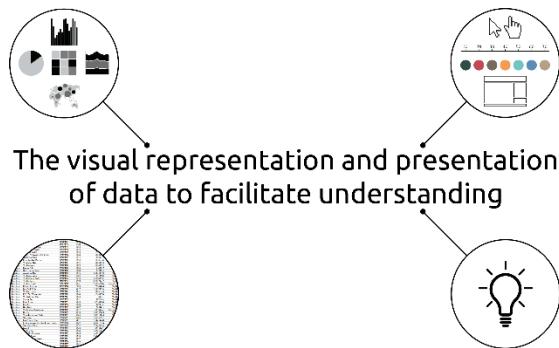


Abb. 6: Definition Daten-Visualisierung nach Kirk (2019)

«Daten-Visualisierung ist die **visuelle Repräsentation** und **Präsentation** von **Daten**, um das **Verständnis** zu erleichtern»

Diese Definition nach Kirk (2019) hebt vier wichtige Bestandteile hervor:

- **Daten**
Auswahl, Bereinigung und analytischer Prozess, um zum visualisierenden Datensatz zu gelangen.
- **Visuelle Repräsentation**
Mit welcher Diagrammart oder Technik werden die Daten präsentiert?
- **Präsentation**
Interaktivität, textliche Anmerkungen, Farben, Komposition der Diagramme/Grafiken, Art der Präsentation.
- **Verständnis**
Dem Betrachter erleichtern die beabsichtigte Botschaft zu erfassen, interpretieren und verstehen.

7.1 Datengetriebener Design Prozess

Der datengetriebene Design Prozess wird im Buch «Data Visualisation – A Handbook for Data Driven Design» von Kirk (2019) im Detail besprochen und ist aus Sicht der Autoren dem versierten Leser sehr zu empfehlen.

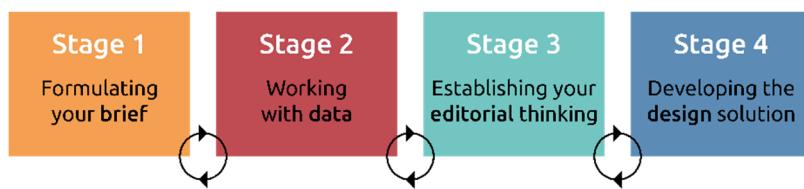


Abb. 7: Stufen des datengetriebenen Design Prozesses, Quelle Kirk (2019)

Die beiden ersten Schritte wurden bereits im vorangehenden Kapitel 6 aus analytischer Sicht erläutert. Neu bezogen auf die vorangehenden Kapitel ist im datengetriebenen Design Prozess nach Kirk (2019) im ersten Schritt die starke Fokussierung auf den End-Betrachter

der Visualisierung und dessen Vorwissen, das Präsentationsmedium, vorhandene Technologien, die Wirkung von Farben, Formen und unterschiedlichen Diagrammen und der Präsentation als Endprodukt. Der Fokus liegt im ganzen Prozess weniger auf den datenanalytischen Teilen, sondern viel mehr auf der Seite der visuellen Kommunikation und dem Design-Prozess eines Grafikers. Kirk (2019) hebt hervor, dass für jede Visualisierung der Prozess ganz durchgeführt werden muss und jedes Endprodukt ein Unikat ist, welches nutzerorientiert und sehr spezifisch entwickelt wurde.

Nach Aigner et al. (2011) ist es für Praktiker jeweils nicht trivial für eine Anwendung eine geeignete Darstellung zu finden. Es sei unumgänglich, sich über repräsentative- und wahrnehmungsbezogene Fragen Gedanken zu machen. Eine praxisorientierte, systematische Betrachtung aus drei Fragen soll Klarheit schaffen:

- i) Was wird präsentiert?
- ii) Warum wird es präsentiert?
- iii) Wie wird es präsentiert?

Für die Beantwortung von i) und ii) müssen die Daten zuerst erforscht und interessante und wichtige Zusammenhänge gefunden werden. Bei Kirk (2019) werden diese zwei Fragen in den ersten drei Schritten aus Abb. 7 behandelt, wobei der dritte Schritt einer der Wichtigsten sei. Beim «Redaktionellem Denken» geht es Kirk darum, ein fundiertes Urteil über den Inhalt der Visualisierung zu bilden. Dieser Schritt entscheidet, was in die Visualisierung aufgenommen wird und unterscheidet nach Ihm die besten und wirkungsvollsten Visualisierungen von den «anderen». Die Fragestellung lautet: «Was zeigt man von all den gegebenen Sachen, die man zeigen kann? »

Ist die Fragestellung des «Was» und «Warum» beantwortet, kommt die eigentliche Visualisierung der Daten und das «Wie», beziehungsweise die Visualisierungstechnik und Präsentation. Kirk (2019) gibt eine sehr breite Übersicht an möglichen Diagrammen und Visualisierungstechniken mit Vor- und Nachteilen wieder. Für das vorliegende Projekt interessante Visualisierungen werden im Anhang Kapitel 9.2 besprochen.

7.2 Narrative Daten-Visualisierungen (Data Storytelling)

«Data Storytelling ist ein strukturierter Ansatz zur Vermittlung von Datenerkenntnissen und umfasst eine Kombination aus drei Schlüsselementen: Daten, visuelle Darstellung und Erzählung. »
(Dykes, 2016)

Narrative (sinnstiftende, erzählende) Datenvisualisierungen sind durch den Begriff «Storytelling» bekannt und zu einem aktuellen Schlagwort in der Datenvisualisierung geworden. Sie werden deshalb nachfolgend kurz erläutert. Oft wird aus den Erkenntnissen einer Datenanalyse eine übergeordnete Geschichte erzählt und der Betrachter wird Schritt für Schritt in die Thematik eingeführt. Oftmals geht man auch von einer Übersicht rein in Details oder führt den Betrachter von einer Erkenntnis zur Nächsten.

Weber (2020) gibt eine gute Einführung in die Thematik und unterscheidet zwischen einem

- Erzählmodus
- Anzeigemodus

Im erzählenden Modus wird die Botschaft, die in den Daten gefunden wird, durch eine vermittelte Instanz kommuniziert, d.h. der Leser erhält die Botschaft erzählt.

Der Anzeigemodus kommt ins Spiel, sobald die Visualisierung den Leser auffordert, innerhalb eines gegebenen Satzes von Optionen wie Zoomen, Filtern oder Auswählen von Objekten zu interagieren, ohne jedoch den narrativen Rahmen zu verlassen. Dieser begrenzte Interaktionsrahmen, in dem der Betrachter mehr spielerische Kontrolle erhält und der Erzähler im Hintergrund bleibt, kann als ein dialogartiger Kommunikationsprozess gesehen werden.

7.3 Feedback mit Ziel Verhaltensänderung

Sollen Visualisierungen erstellt werden, auf welche i) der Endbetrachter einen Einfluss hat und soll ii) mittels der Darstellung dessen Verhalten beeinflusst werden, so bekommt das Feedback eine sehr zentrale und entscheidende Rolle.

Diese Thematik wird separat im Dokument Literaturrecherche «Empirische Studien» behandelt.

7.4 Visualisierungen des Energieverbrauchs

Diese Thematik wird separat im Dokument Literaturrecherche «Empirische Studien» behandelt.

8 State of the Art Tools

Bei der Wahl von geeigneten Tools für die Visualisierung ist grundsätzlich zu unterscheiden ob man sich an Experten richtet, welche die Visualisierungen im Visual Analytics Prozess benötigen um eine besseres Verständnis für die Daten zu erhalten oder ob man sich an End-User richtet welchen man gezielt Information in einem Dashboard, einer Story oder Report bereitstellen will.

Für erstere Gruppe werden heute gängige Skriptsprachen und deren Frameworks verwendet. Gemäß Hayes (2019) ist Python heute mit Abstand am weitesten verbreitet mit den Libraries (Matplotlib, PyData und PlotLy). Ebenfalls wird R mit seinen Libraries oft verwendet.

Blázquez-García et al. (2020) geben folgende Übersicht öffentlich verfügbarer Software im Bereich der Zeitreihenanalyse wieder:

Name	Language	Related research	Code
Point outliers			
tsoutliers	R	Chen and Liu [1993]	https://cran.r-project.org/web/packages/tsoutliers
spirit	Matlab	Papadimitriou et al. [2005]	http://www.cs.cmu.edu/afs/cs/project/spirit-1/www
STORM	Java	Angiulli and Fassetti [2007, 2010]	https://github.com/Waikato/moa/tree/master/moa/src/main/java/moa/clusterers/outliers/Angiulli
SCREEN	Java	Song et al. [2015]	https://github.com/zaqthss/sigmod15-screen
EGADS	Java	Laptev et al. [2015]	https://github.com/yahoo/egads
SCR	Java	Zhang et al. [2016]	https://github.com/zaqthss/sigmod16-scr
libspot	C++	Siffer et al. [2017]	https://github.com/asiffer/libspot
AnomalyDetection	R	Hochenbaum et al. [2017]	https://github.com/twitter/AnomalyDetection
Nupic	Python	Ahmad et al. [2017]	https://github.com/numamenta/nupic
telemanon	Python	Hundman et al. [2018]	https://github.com/khundman/telemanom
OmniAnomaly	Python	Su et al. [2019]	https://github.com/smallcowbaby/OmniAnomaly
OTSAD	R	Carter and Streilein [2012]; Ishimtsev et al. [2017]; Iturria et al. [2020]	https://cran.r-project.org/package=otsad
Subsequence outliers			
tsbitmaps	Python	Kumar et al. [2005]; Wei et al. [2005]	https://github.com/binhmop/tsbitmaps
jmotif	R	Keogh et al. [2005, 2007]; Senin et al. [2015, 2018]	https://github.com/jMotif/jmotif-R
jmotif	Java	Keogh et al. [2005, 2007]; Senin et al. [2015]	https://github.com/jMotif/SAX
saxpy	Python	Keogh et al. [2005, 2007]	https://pypi.org/project/saxpy https://github.com/seninip/saxpy
EBAD	C	Jones et al. [2016]	http://www.merl.com/research/license
GrammarViz	Java	Senin et al. [2015, 2018]	https://github.com/GrammarViz2/grammarviz2_src
Outlier time series			
anomalous	R	Hyndman et al. [2015]	http://github.com/robjhyndman/anomalous-acm

Abb. 8: Zusammenfassung öffentlich verfügbarer Software, Quelle: Blázquez-García et al. (2020)

Kirk (2019) hat auf <https://chartmaker.visualisingdata.com/> eine online-Übersicht, wo man aufgrund des Diagrammtyps mögliche Tools und Bibliotheken vorgeschlagen bekommt. Dies ist für den Experten und Entwickler eine sehr wertvolle Übersicht, weil so viel Zeit gespart werden kann und eine Visualisierung nicht von Grund auf neu entwickelt werden muss.

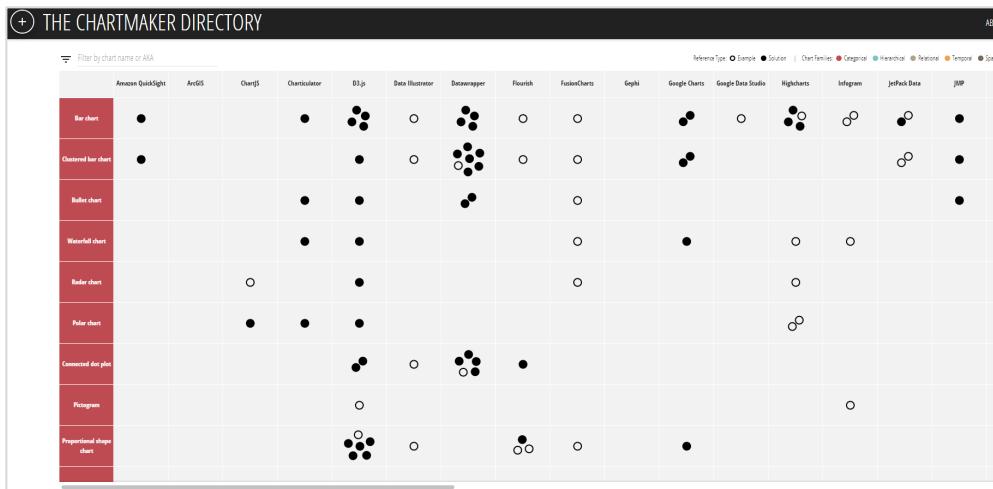


Abb. 9: The Chartmaker Directory, Quelle Kirk (2019) - <https://chartmaker.visualisingdata.com/>

Für Analysten, welche nicht mit Skriptsprachen umgehen möchten, gibt es kostenpflichtige Analysetools, welche den Analyseprozess unterstützen und aussagekräftige Visualisierungen erzeugen können⁵:

- Microsofts Power BI (powerbi.microsoft.com),
- Sisense (sisense.com),
- Looker (looker.com),
- Qualtrics (qualtrics.com),
- Zoho (zoho.com),
- Infragistics (infragistics.com),
- QlikView (qlik.com),
- Tableau (tableau.com),
- SAP Business Objects Lumira (sap.com/swiss/products/lumira.html),
- SAS Business Intelligence (sas.com),
- IBM Cognos (ibm.com/de-de/products/cognos-analytics).

Für die Gruppe der Endkunden gibt es eine Menge an Parametern, welche für die Wahl von geeigneten Tools/Frameworks relevant sind. Zum Beispiel ist relevant für welche Zielplattformen die Visualisierung gemacht werden soll (Web, Mobile, Mikrocontroller, etc.), wie flexibel und konfigurierbar die Dashboards sein müssen, welche Datenmengen erwartet werden und wie viele Benutzer auf eine Plattform zugreifen. Diese Parameter müssen erst geklärt werden, bevor überhaupt eine Aussage über vorhandene Tools gemacht werden kann, da die Anzahl solcher Tools sonst den Rahmen sprengen würden.

⁵ Liste (nicht abschliessend) von financesonline.com

9 Anhang

9.1 Literaturempfehlungen

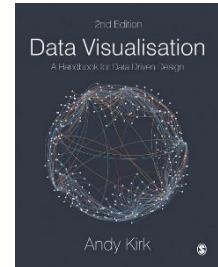
9.1.1 Data Visualisation

Data Visualisation – A Handbook for Data Driven Design

Autor: Andy Kirk

Buchzusammenfassung vom Autor

Mit über 200 Bildern und ausführlichen Beispielen mit Anleitungen und Hinweisen bietet diese neue Ausgabe alles, was Studenten und Wissenschaftler benötigen, um Daten zu verstehen und effektive Datenvisualisierungen zu erstellen. Durch die Kombination von «Wie man denkt» und einer «Wie man produziert»-Mentalität führt dieses Buch den Leser Schritt für Schritt durch die Analyse, Gestaltung und Kuratierung von Informationen zu nützlichen, wirkungsvollen Kommunikationsmitteln.



9.1.2 Visual Analytics

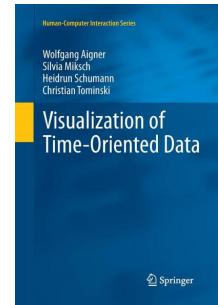
Visualization of Time-Oriented Data

Autoren: Aigner, W., Miksch, S., Schumann, H., Tominski, C.

Buchzusammenfassung von ACM Computing Reviews 16 July 2012

Vier Forscher haben diesen umfangreichen Katalog von Visualisierungstechniken zusammengestellt, bei denen die zeitliche Dimension eine große Rolle spielt.

Die erste Hälfte versucht, die intellektuelle Grundlage hinter den Visualisierungstechniken selbst zu erfassen. Zuerst werden die Visualisierungsaspekte behandelt, dann die Interaktion und die analytische Unterstützung, die die aktuellen Datenvisualisierungstools bieten. In Bezug auf die Datenvisualisierung sind die Techniken um drei einfache Fragen herum strukturiert: Was? Warum? Wie? Die darzustellenden Daten (was) und die zu unterstützenden Benutzeraufgaben (warum) bestimmen den visuellen Darstellungsmechanismus (wie). Dieses Buch enthält auch ein paar Kapitel über Funktionen, die jedes Datenvisualisierungstool unterstützen sollte. Das erste dieser Kapitel, über Interaktion, beschreibt Benutzerabsichten, konzeptionelle und technische Überlegungen, die berücksichtigt werden sollten. Das zweite Kapitel, über die Analytik, berührt lediglich die Oberfläche eines wachsenden Feldes, das eine viel aufwändiger Behandlung verdient. Das Buch enthält Beschreibungen von 101 verschiedenen Möglichkeiten, die zeitliche Dimension von Daten in zwei oder drei Dimensionen darzustellen. Die katalogisierten Techniken beinhalten buchstäblich Dutzende von verschiedenen Möglichkeiten, Zeitreihen darzustellen. Die Darstellungstechniken sind auch online verfügbar: <http://browser.timeviz.net/>

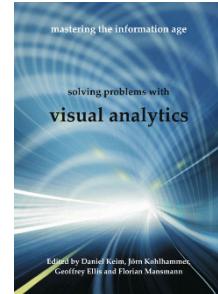


Mastering the Information Age - Solving Problems with Visual Analytics

Autoren: Keim, D., Kohlhammer, J., Ellis, G., Mansmann, F.

Das Buch kann hier kostenlos heruntergeladen werden:

<http://diglib.eg.org/handle/10.2312/14803>



Buchzusammenfassung

Dieses Buch ist das Ergebnis einer Gemeinschaftsarbeit der Partner der von der Europäischen Union finanzierten VisMaster Coordinated Action und fasst den aktuellen Stand der visuellen Analytik über viele Disziplinen zusammen und beschreibt notwendige nächste Schritte in Form einer Forschungs-Roadmap um fortschrittliche Anwendungen zu ermöglichen.

Das erste Kapitel stellt den Problemraum im Hinblick auf die Sinnhaftigkeit sehr grosser, komplexer Datensätze vor und skizziert die Vision für die visuelle Analytik. Das zweite Kapitel befasst sich mit einigen Anwendungsbereichen für die visuelle Analytik und definiert dann die visuelle Analytik im Hinblick auf den Prozess der Wissensentdeckung und berücksichtigt die vielen wissenschaftlichen Disziplinen, die zur visuellen Analytik beitragen. In den Kapiteln 3 bis 8 werden die Arbeiten der spezialisierten Arbeitsgruppen innerhalb des VisMaster-Konsortiums vorgestellt. Jedes dieser Kapitel beginnt mit einem Überblick über den Problembereich und einigen relevanten Hintergrundinformationen. Anschliessend wird ein Überblick über den Stand der Technik in dem jeweiligen Bereich unter Bezugnahme auf die visuelle Analytik gegeben, Herausforderungen und Chancen werden identifiziert und schliesslich Vorschläge, die für das Thema des Kapitels relevant sind, zur Diskussion gestellt. Die übergeordneten Empfehlungen für die Richtung der zukünftigen Forschung in der visuellen Analytik, wie sie von jedem Kapitelautor vorgeschlagen werden, werden im letzten Kapitel zusammengefasst und kategorisiert.

9.2 Visualisierungs-Galerie

Interessante Internet-Seiten mit Visualisierungs-Beispielen:

- <https://docs.aws.amazon.com/quicksight/latest/user/working-with-visual-types.html>
- <https://www.visualisingdata.com/resources/>
- <http://browser.timeviz.net/>
- <https://www.d3-graph-gallery.com/>

Nachfolgend eine Übersicht von Visualisierungen welche für das vorliegende Projekt aus Sicht der Autoren interessant sind.

Quellen: hslu IGE, Aigner et al. (2011) und Kirk (2019).

Übersicht

- 9.2.1 Punktdiagramme
- 9.2.2 Streudiagramme
- 9.2.3 Blasendiagramme
- 9.2.4 Liniendiagramme
- 9.2.5 Flächendiagramme
- 9.2.6 Balkendiagramme
- 9.2.7 Boxplot/Whisker-Plot (Kastengrafik)
- 9.2.8 Wortgrafiken (Sparklines)
- 9.2.9 Sankey-Diagramme
- 9.2.10 Mollier h, x-Diagramme
- 9.2.11 Horizontgraphen
- 9.2.12 Zyklusdiagramme
- 9.2.13 Heatmaps
- 9.2.14 Cluster-Analyse der Tagesverläufe
- 9.2.15 Spiraldarstellung
- 9.2.16 Trend Display
- 9.2.17 Proportionales Symbol-Diagramm
- 9.2.18 Waffeldiagramme
- 9.2.19 Sunburst-Diagramm

9.2.1 Punktdiagramme

BEISPIEL

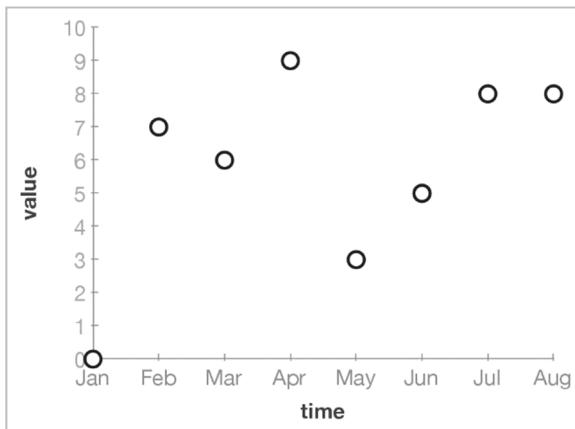


Abb. 10: Punktdiagramm mit einer Variable, Quelle Aigner et al. (2011)

BESCHREIBUNG

Die Daten werden als Punkte auf einem kartesischen Koordinatensystem aufgetragen. Die y-Achse beinhaltet die Variablenwerte, die x-Achse die Zeitkomponente. Für jedes gemessene Zeitwertpaar wird ein Punkt aufgezeichnet.

HINWEISE

- Diese Technik eignet sich besonders gut, um einzelne Werte hervorzuheben.
- Viele Erweiterungen dieser Grundform wie 3D-Techniken (Schichtdarstellung) oder Techniken, die verschiedene Symbole anstelle von Punkten verwenden, sind bekannt.

9.2.2 Streudiagramme

BEISPIELE

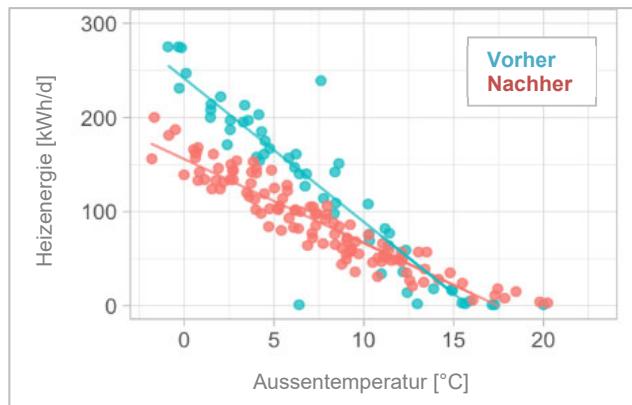


Abb. 11: Streudiagramm Vorher/Nachher, Quelle hslu IGE

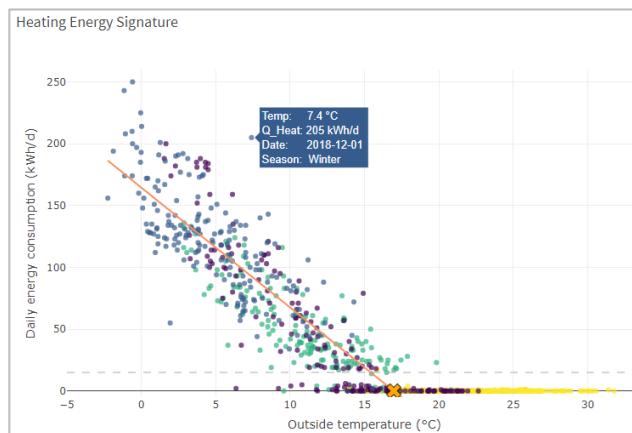


Abb. 12: Streudiagramm Heizsignaturen mit Jahreszeiten, Quelle hslu IGE

BESCHREIBUNG

Im Vergleich zum Punktdiagramm wird hier die Beziehung zweier Variablen zum gleichen Zeitpunkt abgebildet. Der zeitliche Verlauf geht somit verloren.

HINWEISE

- Der Zeitstempel sowie die Zahlenwerte einer Werte-Kombination kann in interaktiven Grafiken als mouse-over Information angezeigt werden.
- Mittels Farben kann der zeitliche Aspekt ebenfalls berücksichtigt werden, einerseits mit einem Farbverlauf oder auch nur mittels unterschiedlicher Farben, welche die Datenpunkte vor und nach einem Datum unterscheiden.
- Referenz- oder Trendlinien können die Interpretation unterstützen.
- Falls Zahlen im Diagramm dargestellt werden müssen, dann sollte man sich auf die Wesentlichsten beschränken, um eine Informationsüberflut zu verhindern.

9.2.3 Blasendiagramme

BEISPIEL

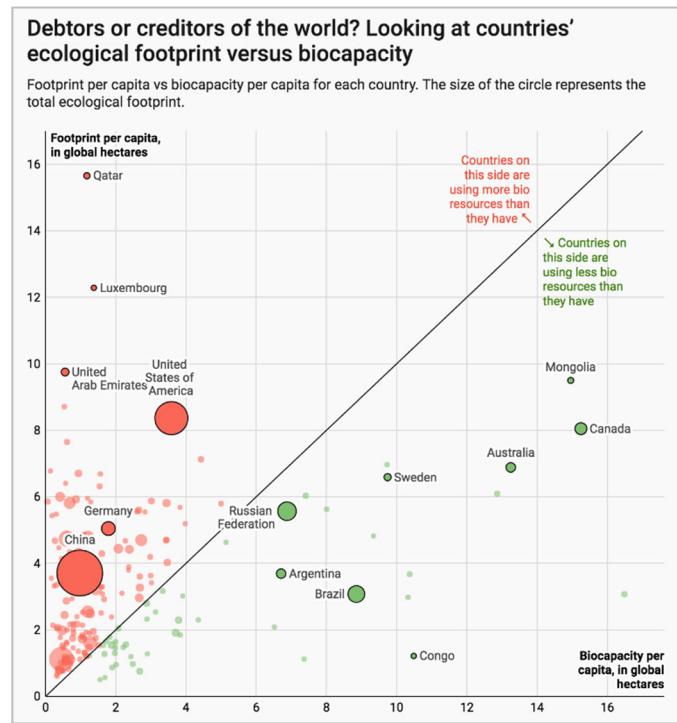


Abb. 13: Blasendiagramm, Kirk (2019)

BESCHREIBUNG

Ein Blasendiagramm zeigt die Beziehungen zwischen drei Variablen. Im Vergleich zum Streudiagramm wird die Grösse der Kreise genutzt, die dritte Variable darzustellen.

HINWEISE

- Interaktive Elemente können hier neben der Anzeige von Werten auch eine Filterung ermöglichen oder Elemente gleicher Kategorien markieren.
- Falls Zahlen im Diagramm dargestellt oder Texte hervorgehoben werden müssen wie in Abb. 13, dann sollte man sich auf die Wesentlichsten beschränken, um eine Informationsüberflut zu verhindern.
- Farben können verwendet werden, um Kategorien zu unterscheiden.
- Grosse Kreise können dahinterliegende kleinere Kreise überlappen. In solchen Fällen kann eine Transparenz benutzt werden.

9.2.4 Liniendiagramme

BEISPIEL

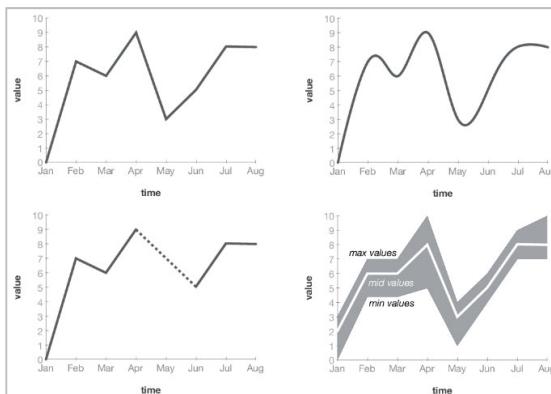


Abb. 14: Liniendiagramme (oben links: gerade Linien, oben rechts Bézier Kurven, unten links: fehlende Daten, unten rechts: Band-Graf), Quelle Aigner et al. (2011)

BESCHREIBUNG

Die wohl gebräuchlichste auf Energiedaten angewendete Darstellungsform. Sie erweitern Punktdiagramme indem sie die Datenpunkte mit Linien verbinden, was ihren zeitlichen Zusammenhang unterstreicht. Folglich konzentrieren sich Liniendiagramme auf die Gesamtform der Daten über die Zeit. Dies ist im Gegensatz zu Punktdiagrammen, bei denen einzelne Datenpunkte hervorgehoben werden.

HINWEISE

- Wie in Abb. 12 dargestellt, können je nach betrachtetem Phänomen unterschiedliche Arten von Verbindungen zwischen den Datenpunkten wie Geraden, Stufenlinien (sofortige Wertänderungen) oder Bézierskurven verwendet werden. Zu beachten ist jedoch, dass man sich nicht in allen Fällen über die Datenwerte im Zeitintervall zwischen zwei Datenpunkten sicher sein kann und dass jede Art von Verbindung zwischen Datenpunkten nur eine Annäherung darstellt.
- Ein weiterer Punkt der Vorsicht ist das Fehlen von Daten. Das blosse Verbinden nachfolgender Datenpunkte kann zu falschen Schlussfolgerungen über die Daten führen. Daher sollte dies für den Betrachter sichtbar gemacht werden, z.B. durch gepunktete Linien
- **Punkt- oder Liniendiagramm?**
 Eine Faustregel nach Dettling (2018) besagt, dass bei Zeitreihendiagramme von Messreihen mit kontinuierlichen Messintervallen die einzelnen Messpunkte mit Linien miteinander verbunden werden sollen. Einzige Ausnahme ist, dass es Lücken bei fehlenden Messwerten geben kann.
- **Seitenverhältnis im Liniendiagramm**
 In einem Papier schlügen Cleveland et al. (1988) die Idee vor, dass die durchschnittliche Neigung der Linien in einem Liniendiagramm 45° betragen sollte. Dies wurde als «Banking to 45° » bezeichnet und hat sich zu einer allgemeinen Regel in zur Bestimmung des idealen Seitenverhältnisses entwickelt. Links ist nach Cleveland die lesbarste.

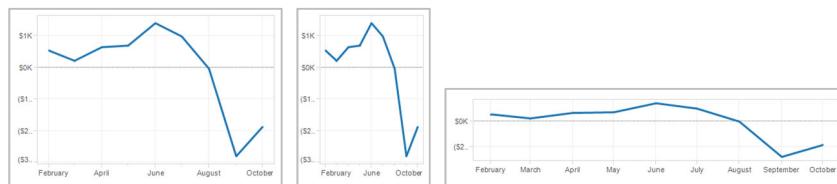


Abb. 15: versch. Seitenverhältnisse einer Zeitreihe. Quelle: eagereyes.org

- Darstellung mehrerer Zeitreihen

Gemäss Dettling (2018) sollen mehrere Zeitreihen normalerweise in verschiedenen Diagrammen dargestellt werden. Diagramme mit zwei oder mehreren y-Achsen sollen die Ausnahme bilden. In beiden Fällen sollten die unterschiedlichen Zeitreihen unterschiedliche Farben bekommen oder durch Interaktivität nur die selektierte farbig sein.

9.2.5 Flächendiagramme

BEISPIEL

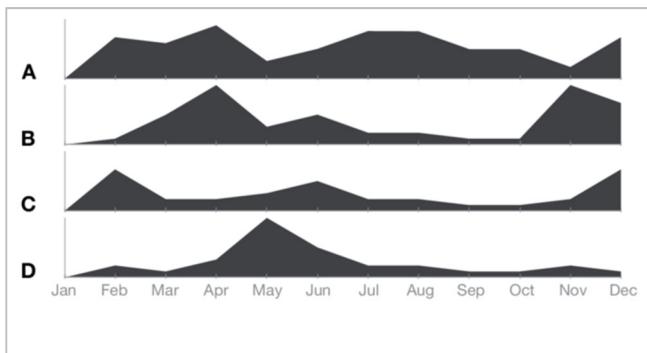


Abb. 16: Flächendiagramme, Quelle Aigner et al. (2011)

BESCHREIBUNG

Flächendiagramme füllen im Vergleich zu Liniendiagrammen den unteren Bereich der gezeichneten Linien aus und verbessern die Wahrnehmung von langen Zeitreihen. Dies ist besonders hilfreich für den Vergleich mehrerer Zeitreihen, welche übereinander gestapelt werden können.

VARIANTEN

Kreisförmiges Flächendiagramm

Um Periodizität in der Zeit zu betonen, ist ein Anwendungsbeispiel die kreisförmige Darstellung der Silhouettendiagramme auf konzentrischen Kreisen.



Abb. 17: kreisförmiges Flächendiagramm, Quelle Aigner et al. (2011)

Gestapelte Flächendiagramme

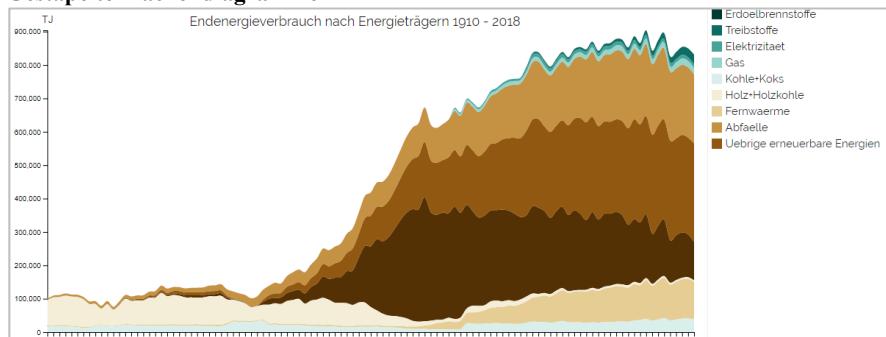


Abb. 18: gestapeltes Flächendiagramm, Datenquelle BfE Gesamtenergiestatistik 2018, Visualisierung <https://streamgraphmaker.guanzo.io/>

Gestapelte Flächendiagramme können für einen Vergleich von Zeitreihen verwendet werden, welche die gleiche Einheit haben und summiert werden können. Bei dieser Art der Darstellung ist Vorsicht geboten, da sie empfindlich auf die Reihenfolge der Schichten reagiert. Nachfolgende Grafiken zeigen, wie die gleichen Daten mit unterschiedlichen Ordnungen das optische Erscheinungsbild der einzelnen Schichten beeinflusst:

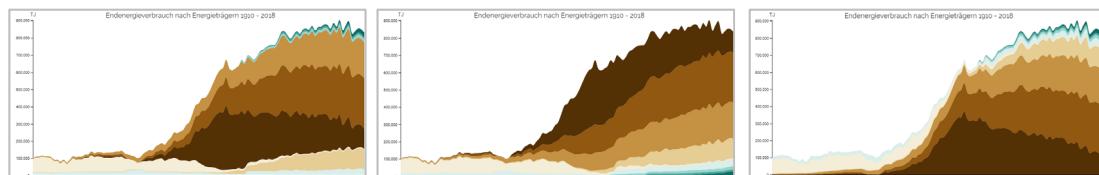


Abb. 19: Vergleich unterschiedlicher Ordnungen, Datenquelle BfE Gesamtenergiestatistik 2018, Visualisierung <https://streamgraphmaker.guanzo.io/>

Ein Vorteil von Layer-Flächendiagrammen ist die Tatsache, dass sie die Gesamtsumme der Werte hervorheben und gleichzeitig Informationen über die Teile liefern, die sie bilden.

Eine interessante Visualisierung ergibt die Anordnung um den Nullpunkt welche die zeitliche Entwicklung besser hervorhebt, jedoch ein direktes Ablesen der Werte nicht mehr zulässt:

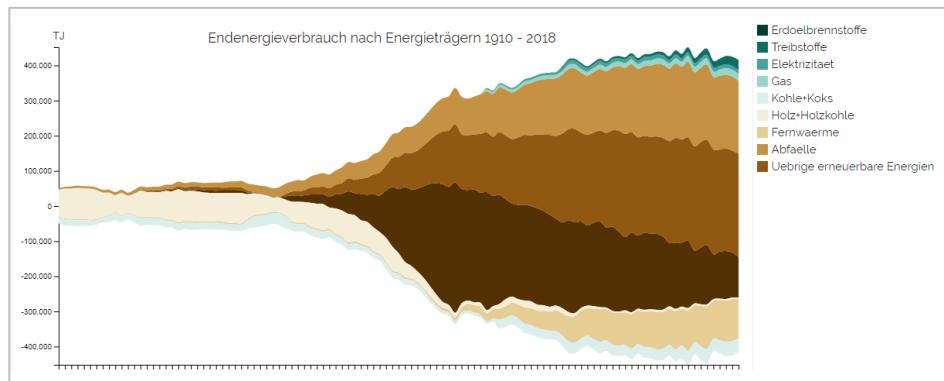


Abb. 20: Stream Graph, Datenquelle BfE Gesamtenergiestatistik 2018, Visualisierung <https://streamgraphmaker.guanzo.io/>

9.2.6 Balkendiagramme

BEISPIEL

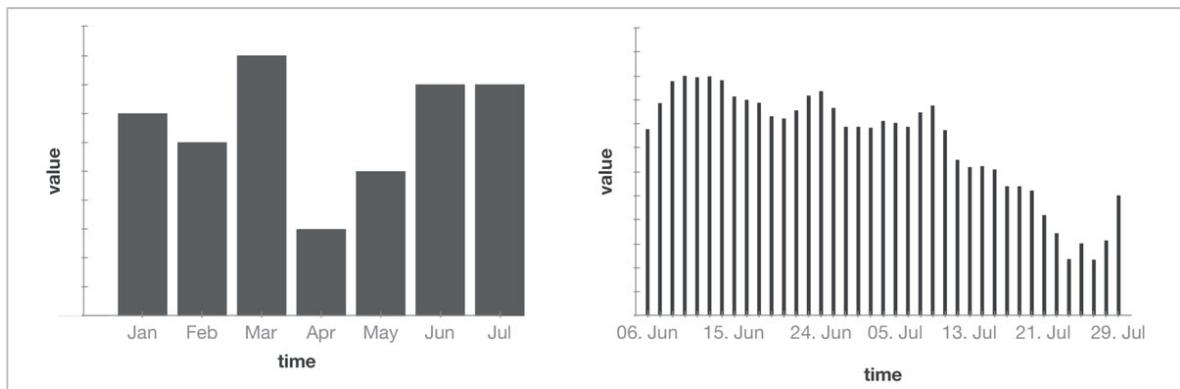


Abb. 21: Balkendiagramme, Quelle Aigner et al. (2011)

BESCHREIBUNG

Balkendiagramme sind eine bekannte und weit verbreitete Darstellungsform, bei der Balken zur Darstellung von Datenwerten verwendet werden. Dies erleichtert den Vergleich im gegenüber zu Punktdiagrammen. Da die Balkenlänge zur Darstellung von Datenwerten verwendet wird, können nur Variablen mit einer Verhältnisskala (mit einer natürlichen Null) dargestellt werden. Im Gegensatz zu Liniendiagrammen betonen Balkendiagramme einzelne Werte wie Punktdiagramme. Eine Variante von Balkendiagrammen, die häufig für die Darstellung grösserer Zeitreihen (z.B. für Stromspitzen) verwendet werden, sind Spike-Diagramme. Wie in Abb. 21 rechts dargestellt, werden die vertikalen Balken so reduziert, dass die Maximum-Tageswerte als Spitzen erscheinen. Auf diese Weise wird ein gutes visuelles Gleichgewicht zwischen der Fokussierung auf Einzelwerte und der Darstellung der Gesamtentwicklung erreicht.

HINWEISE

- Skalenstriche und Rasterlinien können nach Bedarf hinzugefügt werden, um das Ablesen zu vereinfachen.
- Neben den regelmässigen Skalenstrichen kann es gemäss Edward Tufte auch sinnvoll sein, zusätzlich den Minimum- und Maximum-Wert anzugeben da der Betrachter oft nach diesen Werten sucht.
- Werteskala immer mit Null beginnen, um einen fairen visuellen Vergleich zu ermöglichen.
- Wenn möglich nicht bei über jedem Balken den Zahlenwert anzeigen, da dies rasch überladen wirken kann.
- Das Balkendiagramm kann entweder horizontal- oder vertikal angezeigt werden. Es soll die Variante gewählt werden, welche die Daten visuell einfacher und attraktiver präsentiert.
- Ein schmaler aber nicht zu breiter Abstand zwischen den Balken hilft die einzelnen Balken besser zu unterscheiden.
- Je nach Art der Daten bietet sich eine auf- oder absteigende Sortierung der Balken an.

VARIANTEN

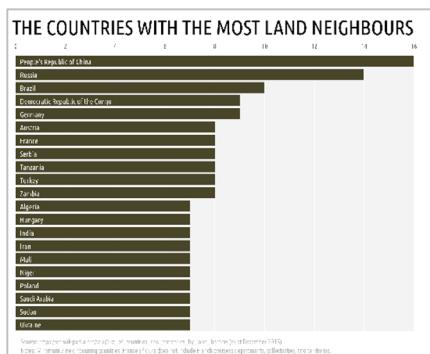


Abb. 22: Balkendiagramm mit Kategorien, Quelle Kirk (2019)

Das nachfolgende Diagramm ermöglicht neben den Primärkategorien den Vergleich von zwei oder mehreren Sekundärkategorien. Oft werden diese farblich unterschiedlich gekennzeichnet.

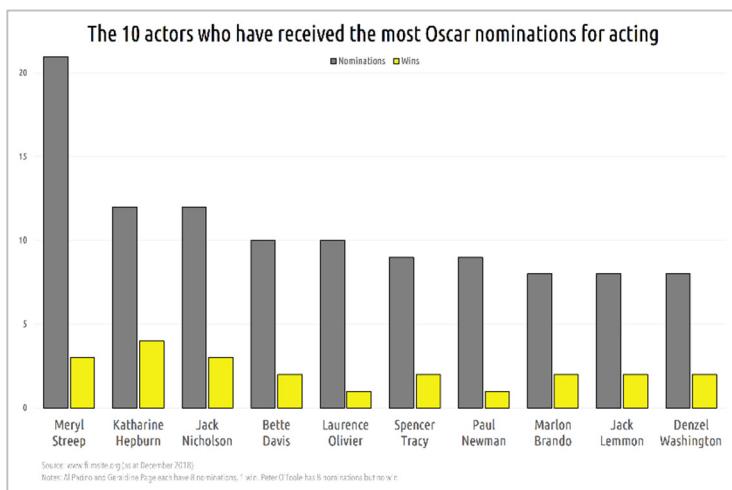


Abb. 23: Gruppiertes Balkendiagramm mit Sekundärkategorien, Quelle Kirk (2019)

Das nachfolgende Aufzählungs-Diagramm hat zusätzlich noch farblich hinterlegte Bänder/Bereiche, welche die Interpretation ggf. vereinfachen können.

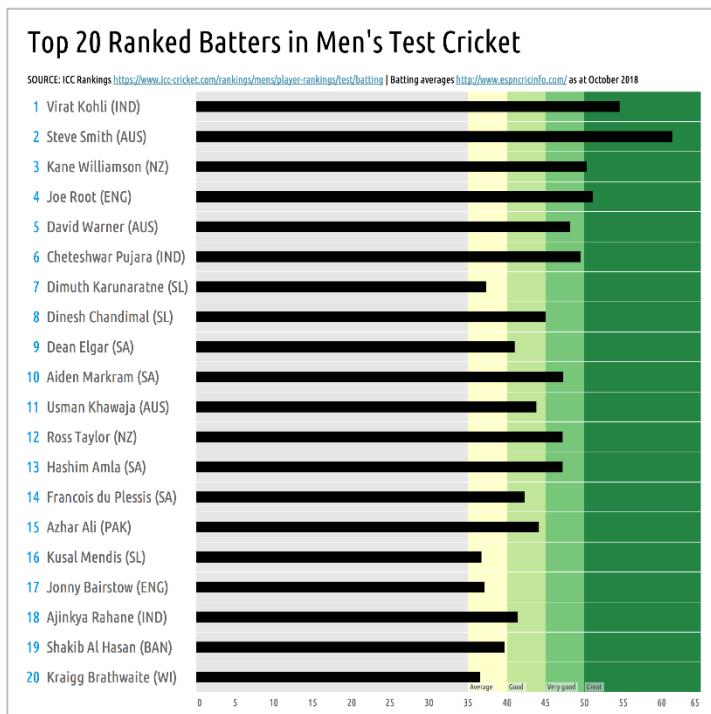


Abb. 24: Aufzählungs-Diagramm, Quelle Kirk (2019)

Im nachfolgenden Kreisdiagramm ist ein spannendes Beispiel eines Balkendiagrammes, welches auf einen Radial übertragen wurde. Jede Stadt wird in einer radialen Darstellung visualisiert und zeigt auf einen Blick den Temperaturverlauf und den Niederschlag.

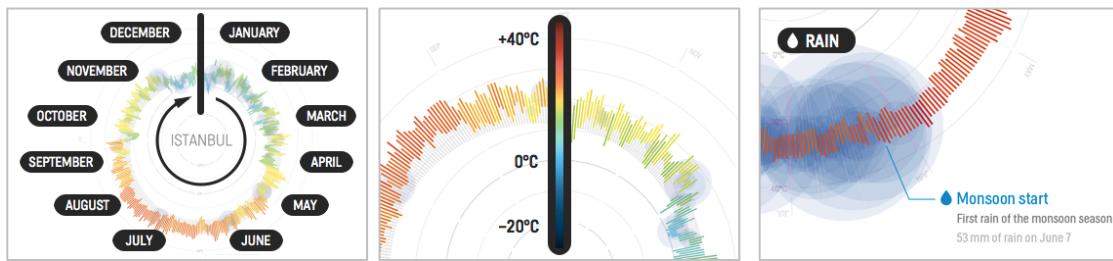


Abb. 25: Aufbau des Wetter-Radials, Quelle weather-radials.com

Ein Wetterradial besteht aus 365 Linien; eine für jeden Tag des Jahres. Die erste ist der 1. Januar ganz oben (12 Uhr), die Tage laufen im Uhrzeigersinn weiter (Abb., links).

Je näher eine Temperaturlinie am Mittelpunkt eines Kreises liegt, desto kälter ist die minimale Temperatur des Tages. Je weiter aussen, desto wärmer ist die Tageshöchsttemperatur. Die Farbe stellt die Tagesmitteltemperatur dar (Abb., Mitte).

Niederschlag (Regen oder Schnee) wird als blauer Kreis dargestellt, der die Menge darstellt (mehr Regen = grösserer Kreis). Die Regenkreise werden in der Mitte der Temperaturlinie des Tages platziert (Abb., rechts).

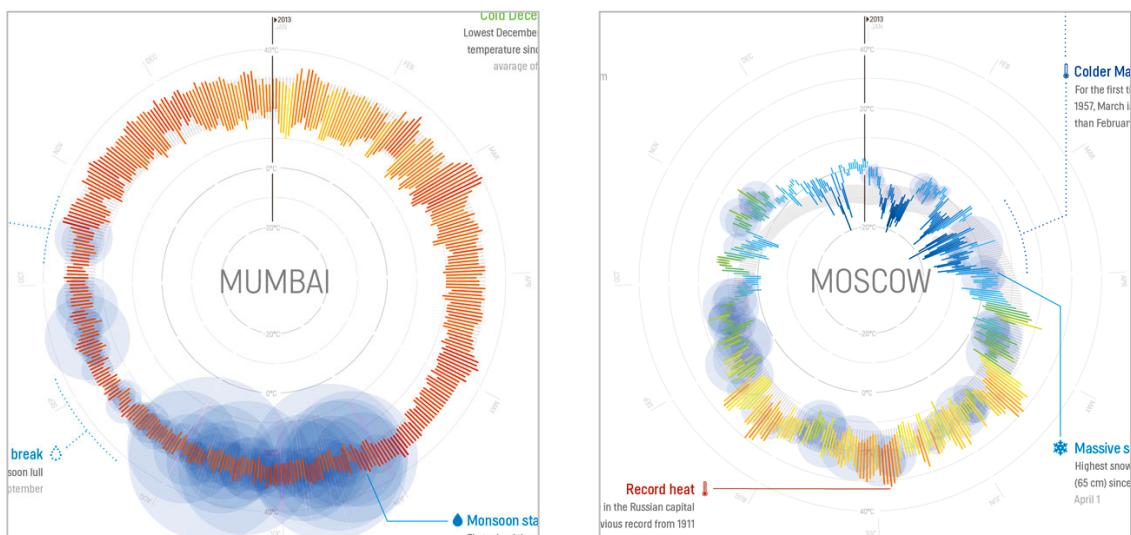


Abb. 26: Exemplarische Wetter-Radiale für Mumbai und Moskau, Quelle weather-radials.com

Das Wasserfall-Diagramm zeigt die Entwicklung eines Ausgangswertes zu einem Endwert:

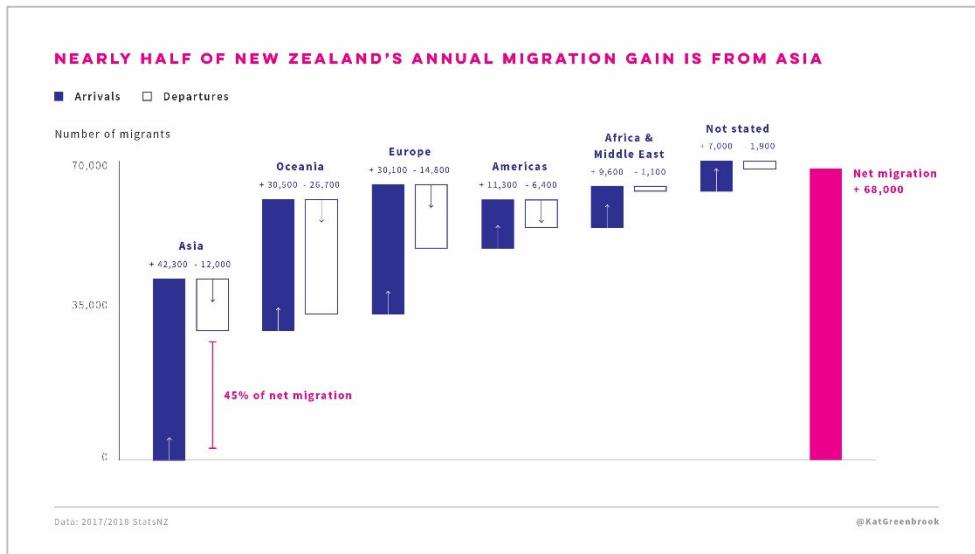


Abb. 27: Wasserfall-Diagramm, Quelle Kirk (2019)

9.2.7 Boxplot/Whisker-Plot (Kastengrafik)

BEISPIELE

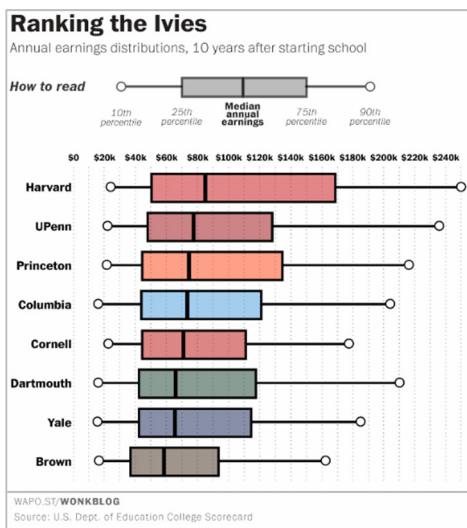


Abb. 28: Boxplot, Quelle Kirk (2019)

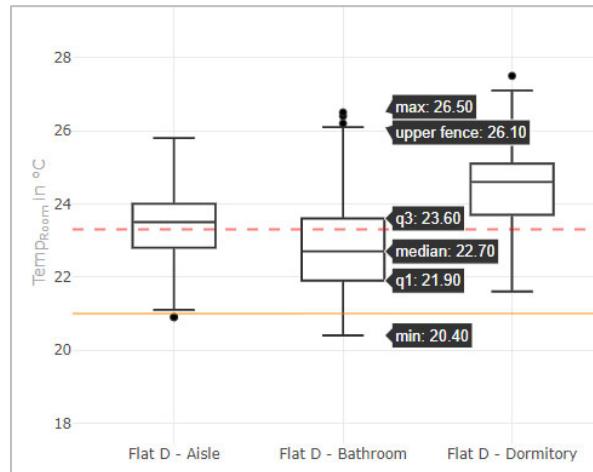


Abb. 29: Boxplots, Quelle hslu IGE

BESCHREIBUNG

Boxplots sind Diagramme, welche zur grafischen Darstellung der Verteilung von Werten verwendet werden. Ein Boxplot fasst dabei verschiedene robuste Streuungs- und Lagemasse in einer Darstellung zusammen. Er soll schnell einen Eindruck darüber vermitteln, in welchem Bereich die Daten liegen und wie sie sich über diesen Bereich verteilen. Deshalb werden alle Werte der sogenannten Fünf-Punkte-Zusammenfassung, also der Median, die zwei Quartile und die beiden Extremwerte (Whisker), dargestellt. Die Extremwerte sind nicht vorgegeben, so können dies die Perzentile 10 und 90 sein oder Minimum- und Maximum-Werte. Optional können noch Ausreißer gekennzeichnet werden.

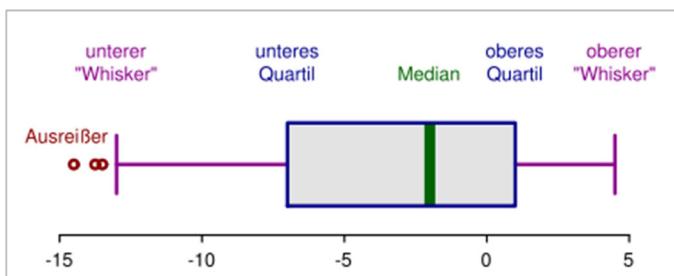


Abb. 30: Erklärungsgrafik gemäß Wikipedia: Box-Plot

HINWEISE

- Mittels Interaktivität können die Werte der Fünf-Punkte-Zusammenfassung eingeblendet werden (siehe Beispiel in Abb. 29).
- In der Visualisierung selbst sollen nur erwähnenswerte Werte als Zahl genannt werden. Ansonsten sollen Achsen-Skalen-Einteilung helfen die Wertebereiche zu identifizieren.
- Die Skalierung der Achse muss nicht bei 0 starten, da es bei dieser Visualisierung darum geht die Verteilung zu zeigen.
- Der Boxplot kann entweder horizontal- oder vertikal angezeigt werden. Es soll die Variante gewählt werden, welche die Daten visuell einfacher und attraktiver präsentiert.

9.2.8 Wortgrafiken (Sparklines)

BEISPIEL



Abb. 31: Wortgrafiken, Quelle Aigner et al. (2011)

BESCHREIBUNG

Edward Tufte beschreibt Wortgrafiken als einfache, wortähnliche Grafiken, die in Text integriert werden sollen. Dies fügt reichere Informationen über die Entwicklung einer Variablen im Laufe der Zeit hinzu, die Worte selbst kaum vermitteln können. Achsen und Bezeichnungen fallen weg, da sie im Kontext des Textes schon bekannt sind. Wortgrafiken lassen sich nahtlos in Textabschnitte integrieren, können als Tabellen angelegt oder für Dashboards verwendet werden. Sie werden zunehmend eingesetzt, um Informationen auf Webseiten (z.B. Nutzungsstatistiken) in Zeitungen (z.B. für Sportstatistiken) oder im Finanzbereich (z.B. für Börsendaten) darzustellen. In der Regel werden miniaturisierte Versionen von Linien- und Balkendiagrammen verwendet. Bei Linienplots können der erste und letzte Wert durch farbige Punkte hervorgehoben und die Werte selbst textuell links und rechts neben der Wortgrafik gedruckt werden (Abb., rot). Darüber hinaus können die Minimal- und Maximalwerte auch durch farbige Punkte gekennzeichnet sein (Abb., blau).

HINWEISE

- Bei Linienplots können der erste und letzte Wert durch farbige Punkte hervorgehoben und die Werte selbst textuell links und rechts neben der Wortgrafik gedruckt werden (Abb., rot).
- Darüber hinaus können die Minimal- und Maximalwerte auch durch farbige Punkte gekennzeichnet sein (Abb., blau).

9.2.9 Sankey-Diagramme

BEISPIEL

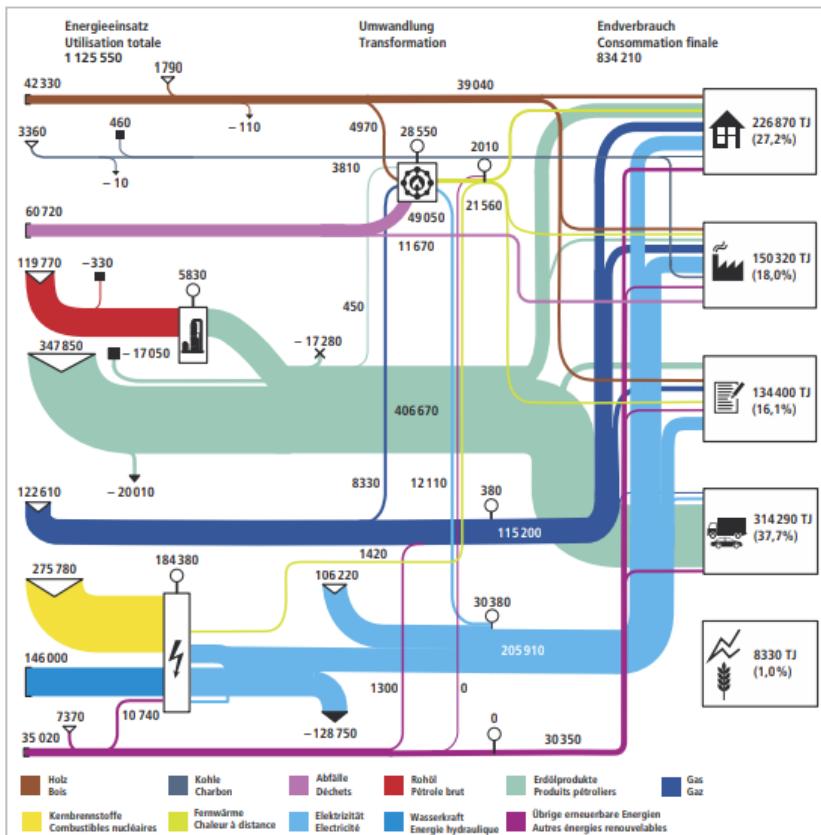


Abb. 32: Energieflussdiagramm der Schweiz 2019, Quelle BfE – Schweizerische Gesamtenergiestatistik

BESCHREIBUNG

Ein Sankey-Diagramm ist eine Visualisierungstechnik, die es ermöglicht, Strömungen darzustellen. Mehrere Entitäten (Knoten) werden verbunden. Deren Breite ist proportional zur dargestellten Menge des Flusses. Es werden in der Regel Mengengrößen abgebildet, die sich auf eine Zeitperiode beziehen.

9.2.10 Mollier h, x-Diagramme

BEISPIEL

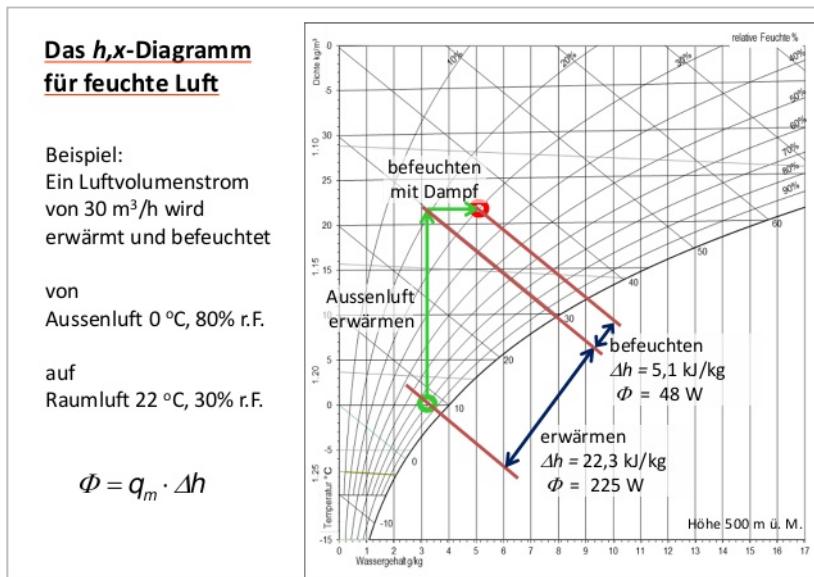


Abb. 33: h, x -Diagramm feuchte Luft, Quelle minergie.ch

BESCHREIBUNG

Das Mollier h,x -Diagramm wurde 1923 von Richard Mollier vorgeschlagen und erlaubt es, Zustandsänderungen feuchter Luft durch Erwärmung, Befeuchtung, Entfeuchtung, Kühlung und Mischung verschiedener Luftpakete zu beschreiben. Es gilt für einen bestimmten Luftdruck (in der Regel für den atmosphärischen Luftdruck), also für isobare Zustandsänderungen. Die Größen Temperatur, Luftfeuchtigkeit, Enthalpie und Dichte können unmittelbar abgelesen werden. Zustandsänderungen können auf grafischem Wege ermittelt werden. Die Grundskala für das h,x -Diagramm ist eine Temperaturskala, die vertikal als y-Achse aufgetragen wird. Die Hilfslinien, die horizontal von links nach rechts gezeichnet werden, sind die "Isothermen", d.h. Linien mit konstanter Lufttemperatur. Während die Isotherme bei 0°C parallel zur horizontalen Achse verläuft, steigen die Isothermen bei höheren Temperaturen aufgrund des Wärmeinhalts des zunehmenden Wassergehalts zunehmend nach rechts an. Die x-Achse stellt den Wassergehalt x bzw. die absolute Feuchte der Luft dar.

9.2.11 Horizontgraphen

BEISPIEL

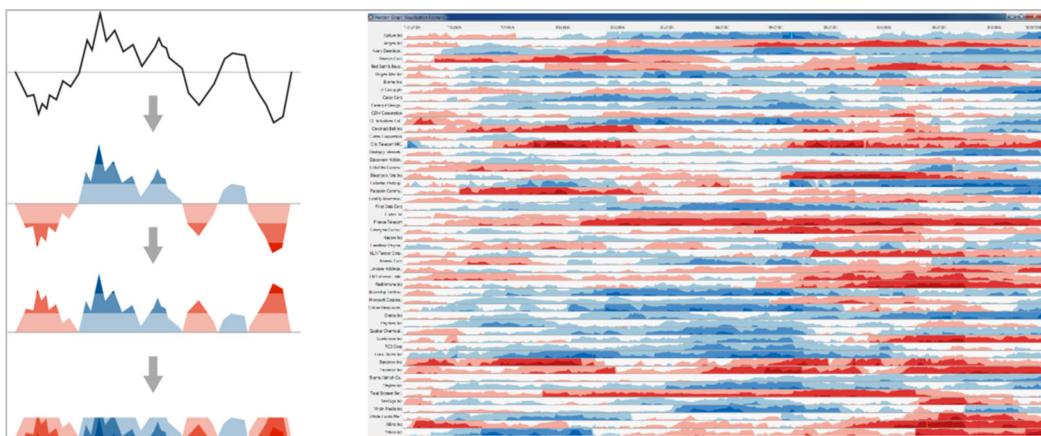


Abb. 34: Horizontgraph, Quelle Aigner et al. (2011)

BESCHREIBUNG

Horizontgraphen sind eine Visualisierungstechnik zum Vergleich einer grossen Anzahl von zeitabhängigen Variablen. Die Horizontgraphen basieren auf der ZweifarbenTechnik. Der linke Teil der Abbildung oben zeigt den Aufbau von Horizontgraphen (von oben nach unten). Ausgehend von einem gemeinsamen Liniendiagramm wird der Wertebereich in gleich grosse Bänder unterteilt, die durch die Erhöhung der Farbintensität in Richtung Maximal- und Minimalwert unterschieden werden, während unterschiedliche Farbtöne für positive und negative Werte verwendet werden. Anschliessend werden negative Werte an der Nulllinie horizontal gespiegelt. Abschliessend werden die Bänder übereinander geschichtet. Auf diese Weise wird weniger vertikaler Raum genutzt, was bedeutet, dass die Datendichte erhöht wird, während die Auflösung erhalten bleibt. Eine Studie habe gezeigt, dass die Spiegelung keine negativen Auswirkungen hat und dass geschichtete Bänder effektiver sind als ein Liniendiagramm.

HINWEISE

- Durch die Komplexität des Aufbaus ist diese Grafik für Layen weniger geeignet.

9.2.12 Zyklusdiagramme

BEISPIEL

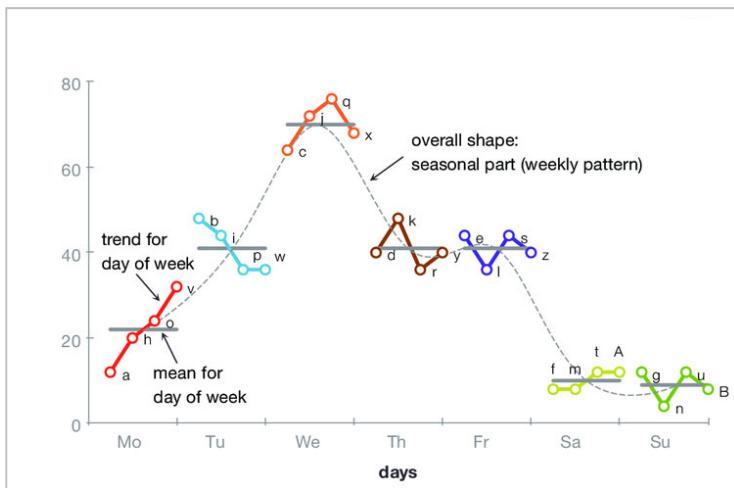


Abb. 35: Zyklusdiagramm, Quelle Aigner et al. (2011)

BESCHREIBUNG

Zyklusdiagramme kommen zum Einsatz, wenn gleichzeitig sowohl Trend- wie auch saisonale Komponenten abgebildet werden sollen. Die x-Achse wird in die saisonale Komponente unterteilt. Die obenstehende Grafik zeigt Wochentage und es wird ein Wochenmuster dargestellt. Die Daten für einen einzelnen Wochentag werden als separates Liniendiagramm dargestellt (Daten des ersten, zweiten, dritten und vierten Montags etc.) sowie der Mittelwert grau für den entsprechenden Tag. So können individuelle Trends für verschiedene Wochentage identifiziert werden. Die Verbindung der grauen Mittelwerte als Liniendiagramm (gestrichelte Linie) welches deutlich ein Spitzenwert am Mittwoch zeigt.

9.2.13 Heatmaps

BEISPIELE

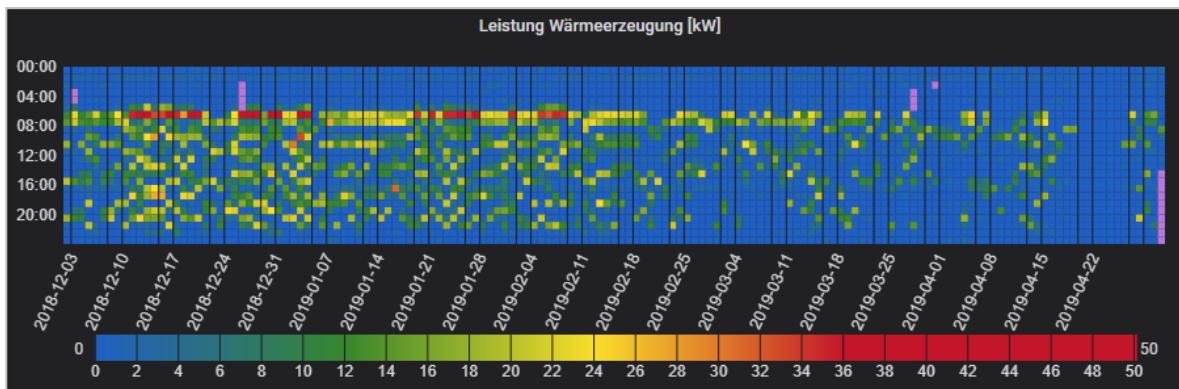


Abb. 36: Heatmap, Quelle hslu IGE

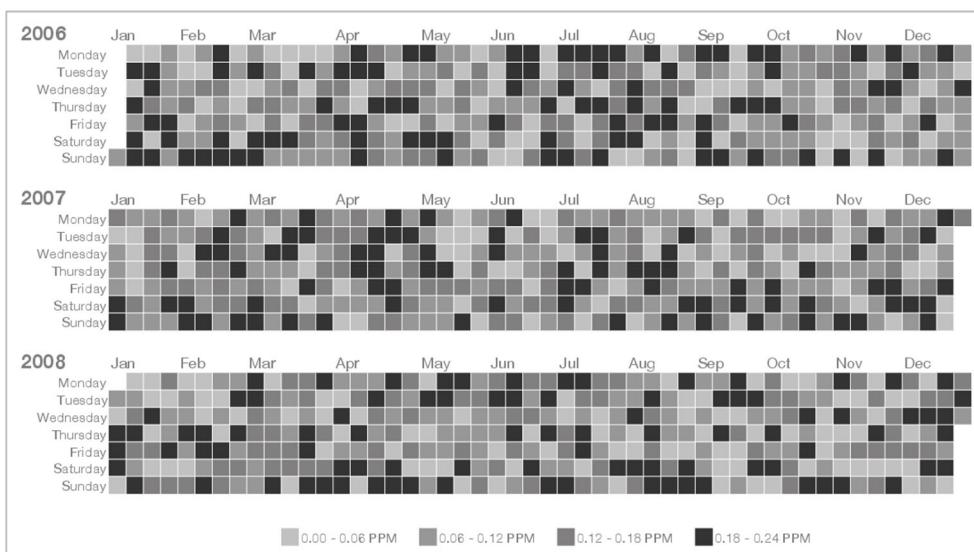


Abb. 37: kalenderbasierte Kachelkarte, Quelle Aigner et al. (2011)

BESCHREIBUNG

Zeitliche Muster können aufzeigen, zu welcher Zeit bestimmte Ressourcen wie stark belastet werden. Heatmaps (Hitzekarten) zeigen quantitative Werte anhand von zwei Achsen auf. Diese können beispielsweise Tag und Stunden, Monat und Tag sein etc. Die Zelle im Schnittpunkt beinhaltet dann den entsprechenden Wert. Dies kann je nach Kontext beispielsweise ein Mittelwert, Maximum/Minimum-Wert sein. Diese Darstellung ermöglicht es den Betrachtern, kurzfristige Muster und langfristige Trends der Daten mit höherer zeitlicher Granularität zu erkennen.

HINWEISE

- Die Zahlenwerte sollen entweder über interaktives mouse-over oder eine klare Legende zur Verfügung gestellt werden.
- Für den Betrachter ist es nicht einfach die Zahlenwerte im Detail zu unterscheiden. Deshalb eignet sich diese Darstellungsart, um eine grobe Übersicht zu erhalten.
- Bei der Wahl der Farbskala (diskret oder kontinuierliche Farben) gibt es kein richtig und falsch. Am besten mehrere Varianten testen und schlussendlich entscheiden, was für den Anwendungsfall optisch und vom Verständnis her am besten wirkt.
- Fehlende Werte können mit einer speziellen Farbe markiert werden.

VARIANTEN



Abb. 38: Kombination verschiedener Aggregationen und Granularitäten, Quelle Aigner et al. (2011)

Die obige Darstellung eignet sich für die Darstellung von Datensätzen vieler Jahre und umfasst in einer Ansicht verschiedene Ebenen der Granularität respektive Aggregationen. Von oben nach unten: Tageswerte, Monatswerte und in der untersten Zeile das Jahrestotal. Eine Spalte entspricht dabei einem Jahr. Fehlende Messdaten werden farblich hervorgehoben.

9.2.14 Cluster-Analyse der Tagesverläufe

BEISPIEL

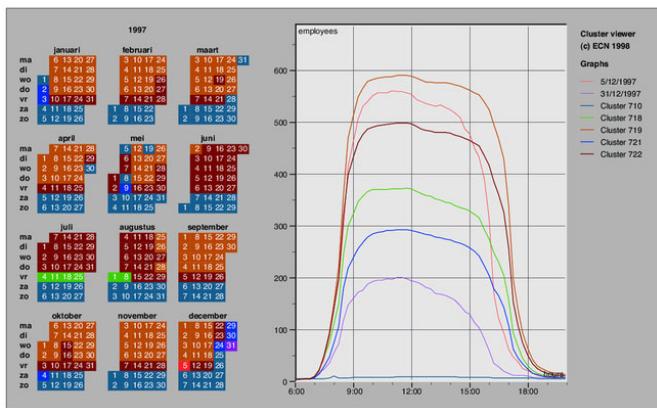


Abb. 39: kalenderbasierte Cluster-Visualisierung, Quelle Aigner et al. (2011)

BESCHREIBUNG

Um bei der Analyse ähnliche und aussergewöhnliche Verläufe hervorzuheben, können tägliche Verläufe in Gruppen zusammengefasst werden. Mittels Cluster-Analyse werden Daten aggregiert und für jedes Cluster mit ähnlichen Verläufen ein Repräsentant dargestellt. Der Repräsentant bildet wieder einen Tagesverlauf (siehe x-y Plot oben rechts).

HINWEISE

- Die Clusterzugehörigkeit einzelner Daten kann je nach Anwendung in einem Kalender farblich kodiert werden.
- Der Benutzer sollte die Anzahl der anzuzeigenden Cluster so einstellen können, dass er den Abstraktionsgrad findet, der den Daten und der jeweiligen Aufgabe entspricht.

9.2.15 Spiraldarstellung

BEISPIEL

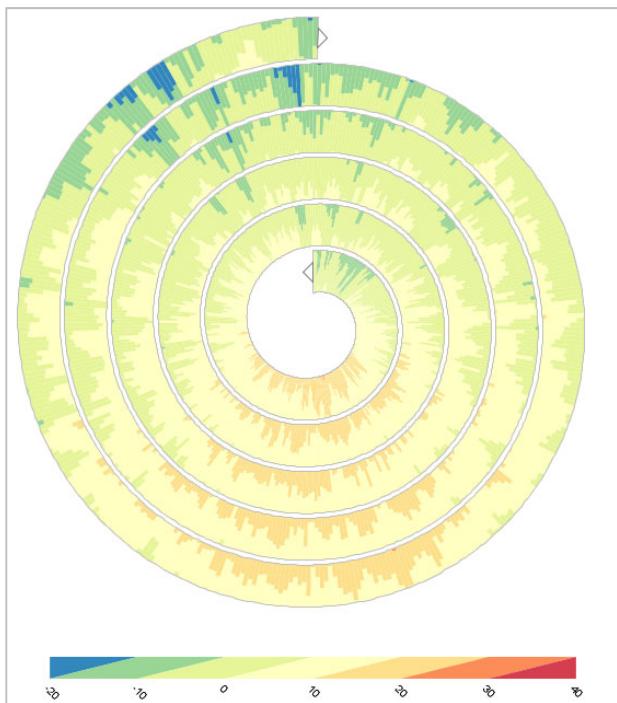


Abb. 40: Spiraldarstellung, Quelle Aigner et al. (2011)

BESCHREIBUNG

Spiraldarstellungen eignen sich bestens für das optische Erkennen von zyklischen Charakteristiken. Die Zeitachse wird dabei als Spirale repräsentiert. Zeitorientierte Daten werden dann dieser Spirale entlang abgebildet. Eine volle Umdrehung kann entweder einen Tag, Monat, Quartal, Jahr, Jahrzehnt etc. abbilden.

HINWEISE

- Abb. 40 zeigt die täglich gemessenen Außentemperaturen von Rostock der Jahre 2006 (ganz innen startend) bis 2010. Die blauen und dunkelgrünen Farben oben links repräsentieren beispielsweise die kalten Winter von 2009 und 2010. Pro Spiralsegment welches einen Tag abbildet, wurden jeweils nur zwei Farben verwendet. Dies macht einerseits die Mustererkennung in der Übersicht einfacher, stellt aber dennoch Detailinformation über die Temperaturhäufigkeit pro Tag dar.

VARIANTEN

Spiraldarstellung zweier Variablen

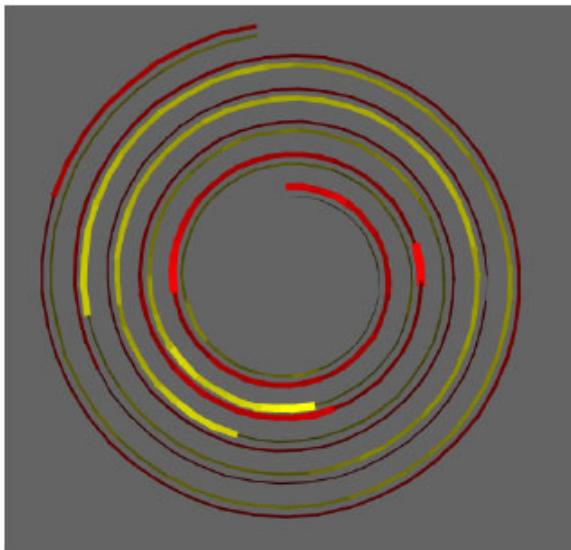


Abb. 41: Spiraldarstellung zweier Variablen, Quelle Aigner et al. (2011)

Diese Darstellung setzt zwei verschiedene Messwerte nebeneinander dar und lässt somit einen saisonalen Vergleich zweier Zeitreihen in einer Grafik zu.

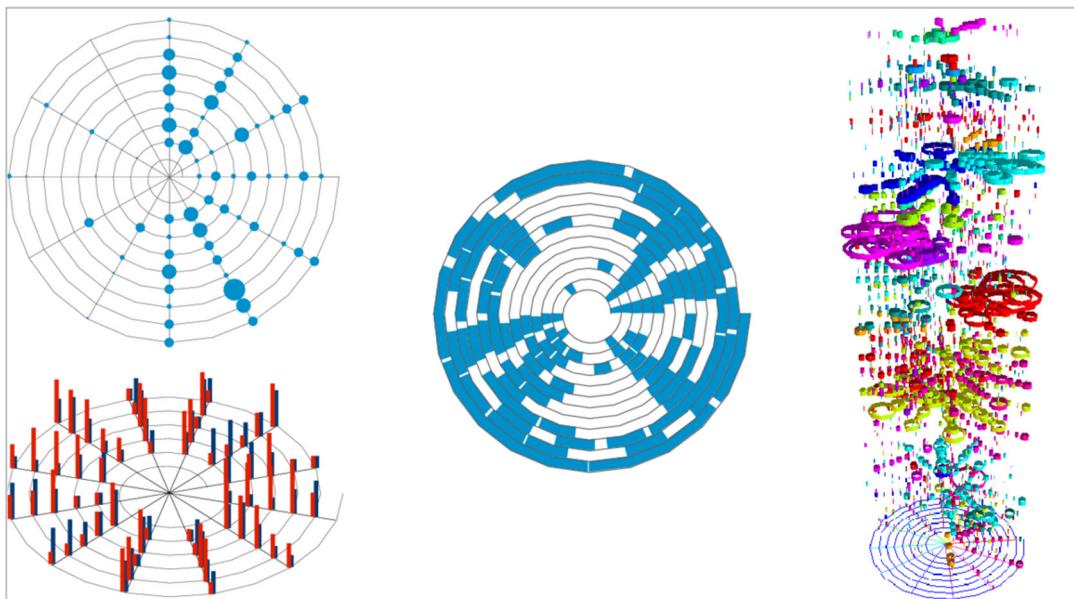


Abb. 42: weitere Beispiele von Spiraldarstellungen, Quelle Aigner et al. (2011)

9.2.16 Trend Display

BEISPIEL

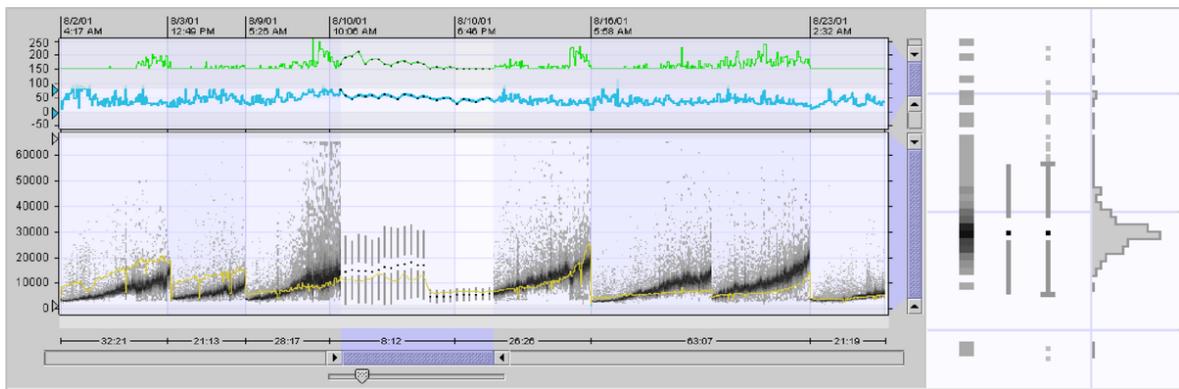


Abb. 43: Trend Display, Quelle Aigner et al. (2011)

BESCHREIBUNG

Die Trend-Display-Technik ermöglicht die Analyse von Trends in grösseren Zeitreihen. Grundsätzlich besteht das Trend-Display-Fenster aus zwei Panels. Das Hauptfenster auf der Unterseite zeigt die gemessenen (Roh-)Daten und das obere Fenster zeigt abgeleitete statistische Werte. Um eine grosse Anzahl von Zeitpunkten zu bewältigen, werden vier verschiedene Detaillierungsstufen verwendet: Dichteverteilungen, Thin-Box-Plots, Box-Plots plus Ausreisser und Balken-Histogramme (von niedrigem bis hohem Detaillierungsgrad). Die verschiedenen Detaillierungsstufen werden je nach verfügbarem Bildschirmplatz automatisch ausgewählt.

HINWEISE

- Die Abb. 43 zeigt eine interaktive Oberfläche für das visuelle Erforschen, welche viel Information übersichtlich darstellt.

9.2.17 Proportionales Symbol-Diagramm

BEISPIEL

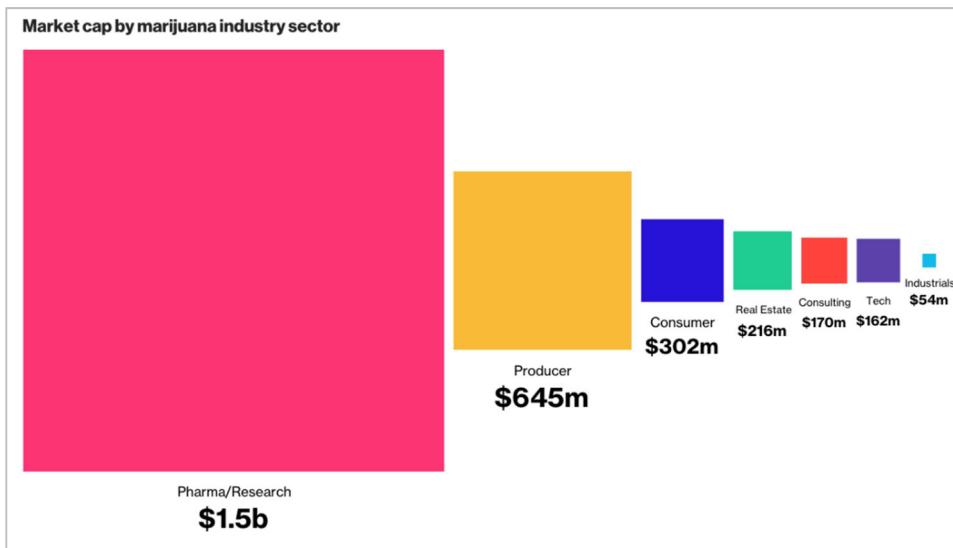


Abb. 44: Proportionales Symbol-Diagramm, Quelle Kirk (2019)

BESCHREIBUNG

Dieses Diagramm zeigt quantitative Werte für verschiedene Kategorien. Als Größenvergleich dient die Fläche.

HINWEISE

- Die farbliche Unterscheidung hilft die Kategorien optisch zu trennen.
- Für den Betrachter kann es schwierig sein, kleine Größenunterschiede zu unterscheiden. Deshalb wirkt diese Visualisierung nur bei grossen Unterschieden der Flächen.
- Bei interaktiven Präsentationen können mit einem mouse-over zusätzliche Informationen eingeblendet werden.
- Falls keine Interaktivität möglich ist, so soll ein Zahlenwert als Vergleich angebracht werden.
- Der optische Vergleich soll über die Fläche stattfinden und nicht über die Seitenlänge der Quadrate. Werden Kreise dargestellt, so soll ebenfalls die Fläche und nicht der Radius als Größenvergleichseinheit gebraucht werden.
- Die Reihenfolge sollte, wenn möglich der Grösse nach sein oder einer logischen, aus dem Kontext herauskommenden Folge.

9.2.18 Waffeldiagramme

BEISPIEL

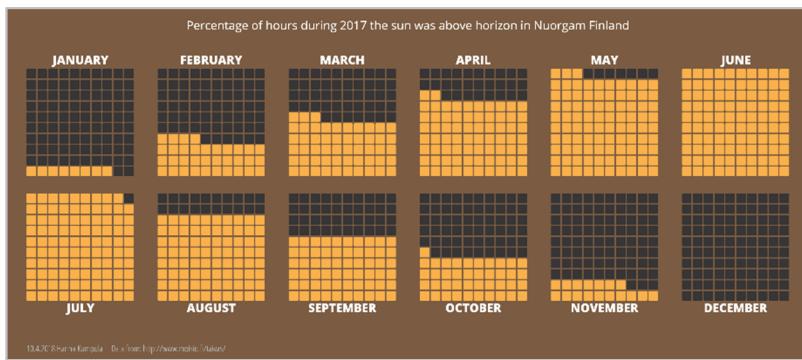


Abb. 45: Waffeldiagramm, Quelle Kirk (2019)

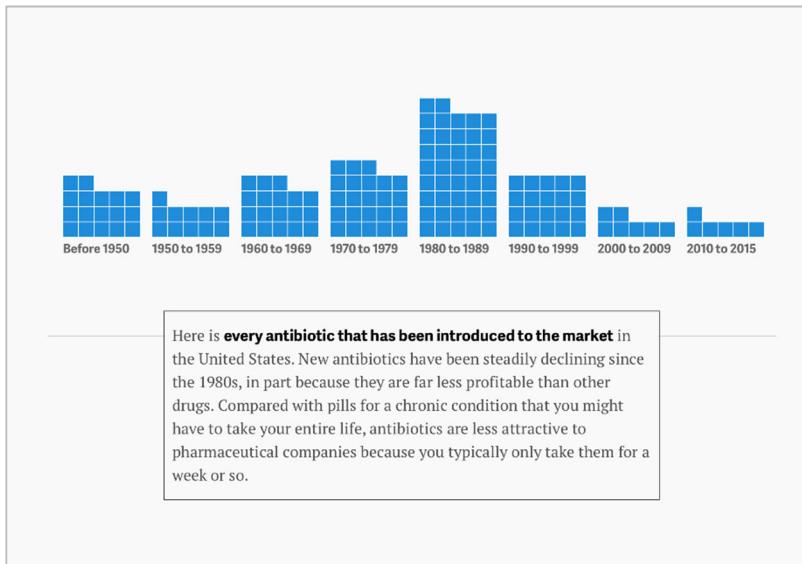


Abb. 46: Waffeldiagramm, Quelle datavizproject.com

BESCHREIBUNG

Ein Waffeldiagramm zeigt prozentuale Werte an wie beispielsweise ein Kreisdiagramm, jedoch einfacher mit Quadraten. Kleine Quadrate, meistens hundert, werden in einem Gitternetz angeordnet. Über Farbkodierung wird der prozentuale Zahlenwert dargestellt. Es kann dazu gebraucht werden einen Fortschritt anzudeuten. Wie Abb. 45 zeigt, können mehrere Waffeldiagramme nebeneinander zusammengesetzt werden, um einen Vergleich zwischen verschiedenen Diagrammen zu zeigen. Das Waffeldiagramm ist einfacher zu lesen als ein Kreisdiagramm, weil hier keine Winkel miteinander verglichen werden müssen und dies über die Höhe geschieht.

9.2.19 Sunburst-Diagramm

BEISPIEL



Abb. 47: Sunburst-Diagramm, Quelle Kirk (2019)

BESCHREIBUNG

Ein Sunburst-Diagramm (Sonnenausbruch) zeigt Werte von Kategorien hierarchischer Beziehungen über mehrere Ebenen. Im Zentrum ist die Hauptebene und je weiter aussen desto detaillierter die Aufsplittung.

HINWEISE

- Farben werden oft gebraucht, um Kategorien noch besser hervor zu heben.
- Mittels Interaktivität können die Zahlenwerte eingeblendet werden.
- Falls Zahlen im Diagramm dargestellt werden müssen, dann sollte man sich auf die Wesentlichsten beschränken, um eine Informationsüberflut zu verhindern.
- Die Reihenfolge sollte, wenn möglich einer logischen, aus dem Kontext herauskommenden Folge nach sein.

10 Literaturverzeichnis

- Adhikari, A., DeNero, J., 'The Foundations of Data Science'.
<https://www.inferentialthinking.com/chapters/intro>
- Andrienko, N., Andrienko, G., (2006). 'Exploratory Analysis of Spatial and Temporal Data', Springer, <https://doi.org/10.1007/3-540-31190-4>
- Aggarwal, Charu C. (2015): Data Mining. Cham: Springer International Publishing.
- Bertini, E., Lalanne, D., (2010). 'Investigating and reflecting on the integration of automatic data analysis and visualization in knowledge discovery', SIGKDD Explorations, 11(2): pp. 9–18
- Blázquez-García, Ane; Conde, Angel; Mori, Usue; Lozano, Jose, (2020): 'A review on outlier/anomaly detection in time series data'.
- Brick Schema (2020). <http://brickschema.org/>
- Card, S. K., Mackinlay, J. (1997). 'The structure of the information visualization design space'. Proceedings of the IEEE Symposium on Information Visualization (InfoVis '97), pp. 92–99.
- Chandola, Varun; Banerjee, Arindam; Kumar, Vipin, (2009). 'Anomaly detection: A Survey', ACM Comput. Surv. 41 (3), pp. 1–58.
- Cleveland, William S., McGill, Marylyn E., McGill, Robert, (1988). 'The Shape Parameter of a Two-Variable Graph'. Journal of the American Statistical Association Vol. 83, No. 402 pp. 289-300 (12 pages)
- Dettling, M. (2018), 'Applied Time Series Analysis'. Institute for Data Analysis and Process Design ETH Zürich.
- Donoho, D. (2015). '50 years of Data Science'. Princeton NJ, Tukey Centennial Workshop. <http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf>
- Dykes, B. (2016). Data storytelling: The essential data science skill everyone needs. Forbes. Retrieved August 26, 2018 from <https://www.forbes.com/sites/brentdykes/2016/03/31/data-storytelling-the-essential-data-scienceskill-everyone-needs/#6126f53e52ad>
- Aigner et al., (2011). 'Visualization of Time-Oriented Data', Springer
<https://doi.org/10.1007/978-0-85729-079-3>, <http://browser.timeviz.net/>
- Fercu, M., Kistler, R., Egli, A., Gallati, J., (2010). 'MEGA; Mehr Energieeffizienz durch gezielte Anwender-Informationen – Schlussbericht', Hochschule Luzern - Technik und Architektur.
- Gupta, Manish; Gao, Jing; Aggarwal, Charu; Han, Jiawei (2014). 'Outlier Detection for Temporal Data', Synthesis Lectures on Data Mining and Knowledge Discovery 5 (1), pp. 1–129. DOI: 10.2200/S00573ED1V01Y201403DMK008.
- Hawkins, D. M., (1980). 'Identification of outliers'. Springer Netherlands, New York.
- Hayes, B., (2019). 'Programming Languages Most Used and Recommended by Data Scientistsin', Business over Broadway
URL: <http://businessoverbroadway.com/2019/01/13/> [19.03.2019]
- Keim et al., (2010). 'Mastering the Information Age - Solving Problems with Visual Analytics', Eurographics Association,
https://kops.uni-konstanz.de/bitstream/handle/123456789/12737/VisMaster-Book_127373.pdf?sequence=2

- Kirk, A., (2020). ‘Data Visualisation – A Handbook for Data Driven Design’. SAGE Publications London. <https://book.visualisingdata.com>
- Niebler, P., Lindner, D. (2018). ‘Datenbasiert entscheiden’. Springer Fachmedien Wiesbaden GmbH, <https://doi.org/10.1007/978-3-658-23928-2>
- Nielsen, A., (2019). ‘Practical Time Series Analysis’. O’Reilly Media Inc.
- Project Haystack (2020). <https://project-haystack.org/>
- Schwarz, J. (2013). ‘Forschungsmethoden - Zeitreihenanalyse’. Manuskript der Vorlesung MSc Banking & Finance, Hochschule Luzern.
- Shneiderman, B., (1996). ‘The eyes have it: A task by data type taxonomy for information visualizations’. IEEE Symposium on Visual Languages, pp. 336–343.
- Thomas, J., Cook, K., (2005). ‘Illuminating the Path: Research and Development Agenda for Visual Analytics’. IEEE-Press
- Tukey, J. W. (1962). ‘The future of data analysis’. The Annals of Mathematical Statistics, 33(1), pp. 1–67. <http://projecteuclid.org/euclid.aoms/1177704711>
- Tukey, J. W. (1977). ‘Exploratory Data Analysis’. Addison Wesley Publishing Company
- Weber, W., ‘Exploring narrativity in data visualization in journalism’. Data Visualization in Society, Engebretsen et. al. (2020), doi 10.5117/9789463722902_ch18 https://digitalcollection.zhaw.ch/bitstream/11475/19886/2/2020_Weber_Narrativity-data-visualization-journalism.pdf
- Wegner, P. (1997). ‘Why Interaction Is More Powerful Than Algorithms’. Communications of the ACM, 40(5) pp. 80–91.