Summary

Jörg Cassens

Data and Process Visualization
SoSe 2017

Inhaltsverzeichnis

1 Introduction 1

2 History 2

3 Semiotics 3
3.1 Communication . . . 3
3.2 Semiotics . . . 4
3.3 Classification Framework . . . 6

4 Perception 7
4.1 Physiology . . . 7
4.2 Color . . . 10
4.3 Processing Pipeline . . . 12
4.4 Attention . . . 13

5 Data 16
5.1 Communicate . . . 16
5.2 Process . . . 18
5.3 Aspects of Data . . . 18
5.4 Data Preparation . . . 21
5.5 Focus . . . 24

6 Representation 26
6.1 Components . . . 26
6.2 Placement . . . 28
6.3 Method . . . 32

7 Presentation 34
7.1 Color . . . 35
7.2 Readability . . . 36
7.3 Interactivity . . . 38
7.4 Annotation . . . 39
7.5 Math . . . 41

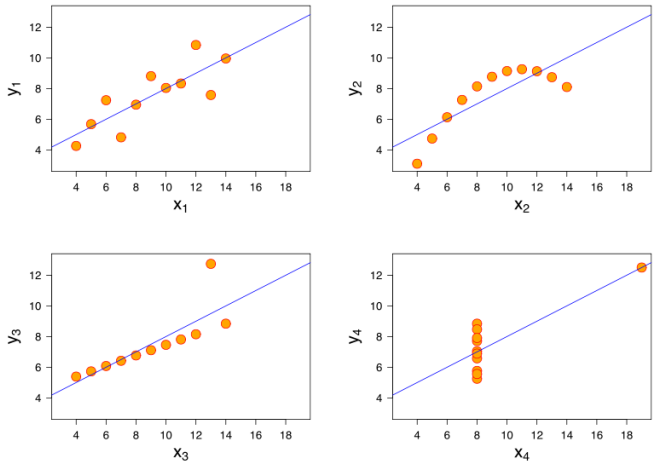
1 Introduction

Data?

x1	y1	x2	y2	x3	y3	x4	y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

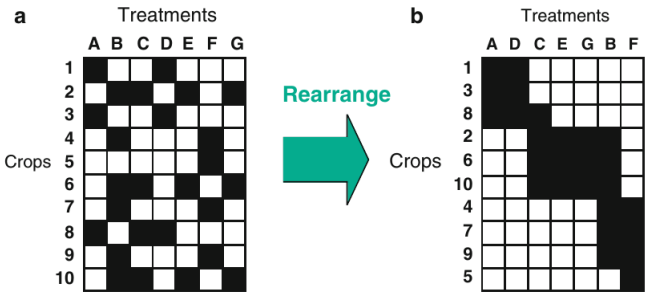
Sources: Anscombe, "Graphs in Statistical Analysis", as cited by Kirk (2012); Tufte (2001)

Insight



Sources: Anscombe, "Graphs in Statistical Analysis", as cited by Kirk (2012); Tufte (2001)

Effort



Reordering gives insight (Spence, 2014)

Definition

"visualization: the activity of forming a mental model of something" (Spence, 2014)

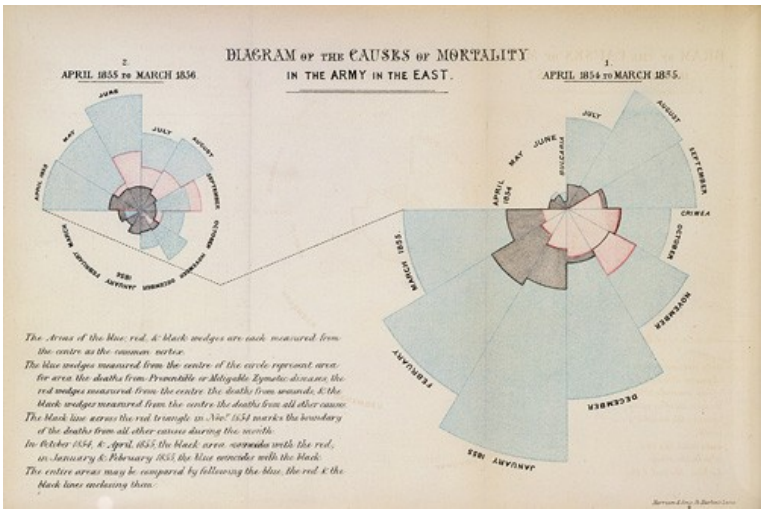
- Visualization is then, by definition, a human activity
  - Nevertheless, it can be enhanced immensely by means of computers
- What is your definition?

2 History

Snow

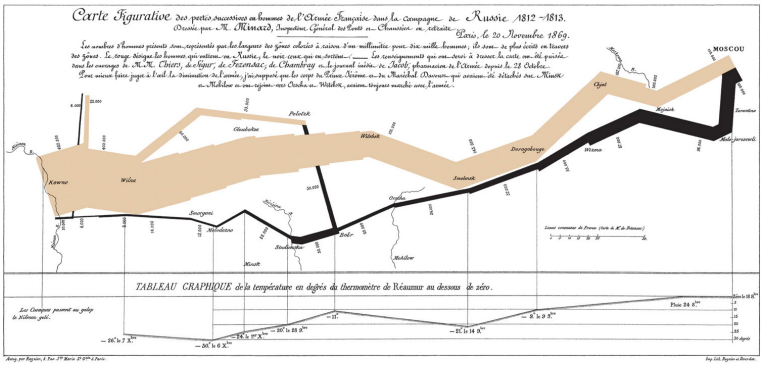
John Snow – Cholera Epidemic of London (Detail) (Spence, 2014)

Nightingale



Florence Nightingale – Cause of death over time (Source: Jänicke (2016))

Minard



Charles Minard – Napoleon's Russian Campaign

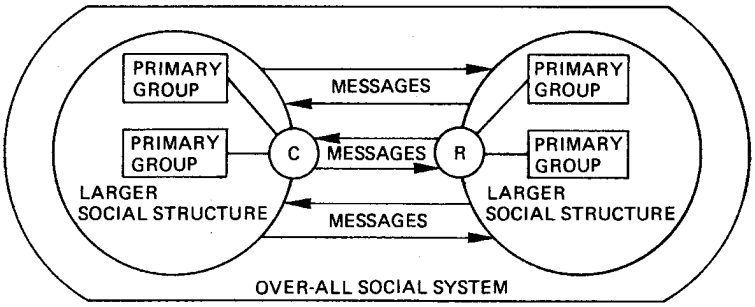
3 Semiotics

3.1 Communication

Communication: Kirk

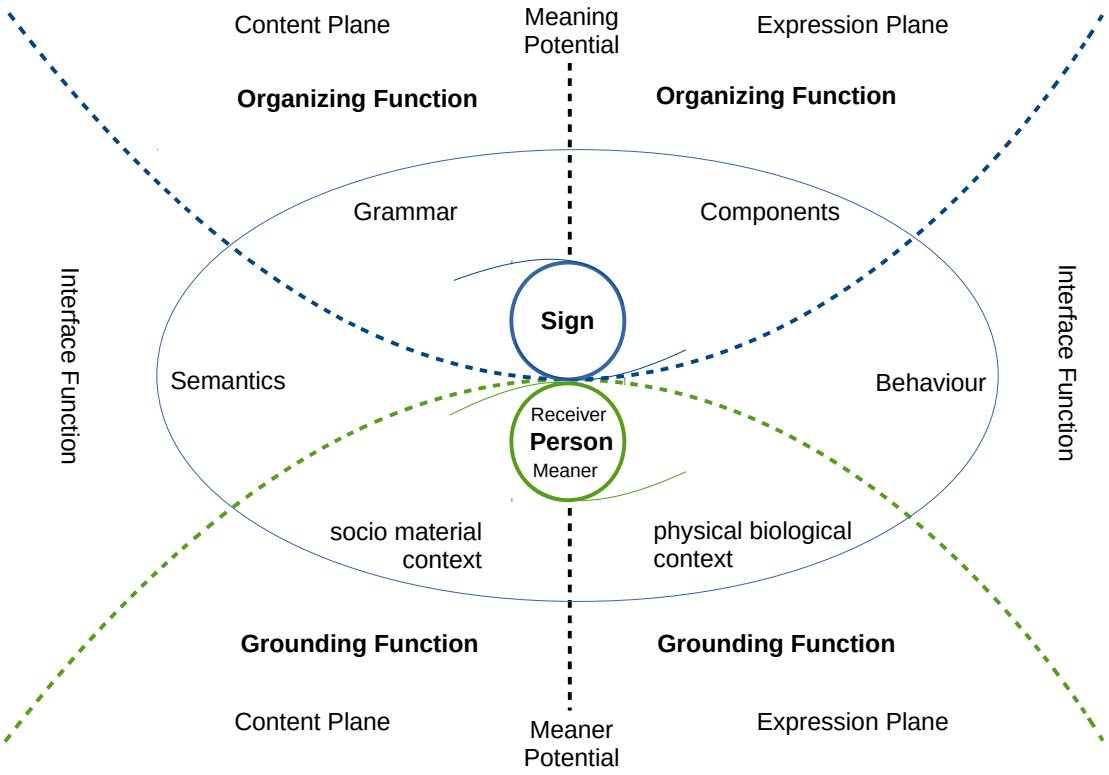
Source: Kirk (2012)

Communication: Riley & Riley



Source: Riley & Riley, here: Kress and van Leeuwen (2006)

Communication: Wegener



Source: Wegener (2011, 2015)

3.2 Semiotics

Sign

- The key notion in any semiotics is the 'sign'
- *Different starting point: not descriptive like Peirce, but functional and social*
- Example drawing was made by a 3-year-old boy
  - Sitting on his father's lap, he talked about the drawing as he was doing it
  - ''Do you want to watch me? I'll make a car . . . got two wheels . . . and two wheels at the back . . . and two wheels here . . . that's a funny wheel . . . ''
  - When he had finished, he said, ''This is a car.''
- This was the first time he had named a drawing, and at first the name was puzzling
- How was this a car?
- He had provided the key himself: 'Here's a wheel.'

(Kress and van Leeuwen, 2006)

A Car

Source: Kress and van Leeuwen (2006)

Car-ness

- A car, for him, was defined by the criterial characteristic 'has wheels', and his representation focused on this aspect
- What he represented was, in fact, 'wheelness'
- Wheels are a plausible criterion to choose for 3-year-olds, and the wheel's action, on toy cars as on real cars, is a readily noticed and describable feature

- This boy's interest in cars was, for him, most plausibly condensed into and expressed as an interest in wheels
  - Choosing what to represent (''the signified'')
- Wheels, in turn, are most plausibly represented by circles
  - Choosing how to represent (''the signifier'')

(Kress and van Leeuwen, 2006)

Representation

> Shortened version: ''We see representation as a *process* in which the makers of signs (. . .) seek to make a *representation of some object or entity,* whether physical or semiotic, and in which *their interest* in the object (. . .) is (. . .) arising out of the cultural, social and psychological *history of the sign-maker,* and focused by the *specific context* in which the signmaker produces the sign. That '*interest*' is the source of the *selection* of what is seen as the *criterial aspect of the object,* and this *criterial aspect* is then regarded as *adequately representative* of the object in a given context. In other words, it is never the 'whole object' but only ever its *criterial aspects* which are *represented.*'' Kress and van Leeuwen (2006)

Sign-Making

- The criterial aspects are represented in what seems to the sign-maker the most apt and plausible fashion, and the most apt and plausible representational mode
- Sign-makers thus 'have' a meaning, the signified, which they wish to express, and then express it through the semiotic mode(s) that make(s) available the subjectively felt, most plausible, most apt form, as the signifier
- This means that in social semiotics the sign is not the pre-existing conjunction of a signifier and a signified, a ready-made sign to be recognized, chosen and used as it is
- We see signs as motivated – not as arbitrary – conjunctions of signifiers (forms) and signifieds (meanings)
- Signs are never arbitrary, and 'motivation' should be formulated in relation to the sign-maker and the context in which the sign is produced,

(Kress and van Leeuwen, 2006)

3.3 Classification Framework

Classification

- Starting with the types of questions users have, the framework supports the selection of data mining and visualization work flows as well as deployment options that answer these user questions.
- We look at the following aspects
  - Level of analysis
  - Types of analysis
  - Intended audience (and/or producer)
  - Medium used
- Some projects aim to answer more than one question

Level of Analysis

- ○ Micro level, or the individual level
  - Small data sets, typically between 1 and 100 records
  - e.g. a person and his friends
- ◐ Meso or the group level
  - About 101 to 10,000 records
  - e.g. researchers at a single university
- ● Macro, global or population level
  - Typically exceeding 10,000 records
  - e.g. pertaining an entire country

Types of Analysis

- 🛈 Statistical Analysis/Profiling
  - What are the entities that are being described (e.g. persons, grants, publications)?
- 🕒 Temporal Analysis: *When*
  - Does the visualization show a development over time?
- 🌍 Geospatial Analysis: *Where*
  - Does the visualization include information about location?
- ≡ Topical Analysis: *What*
  - What is the topical area of the visualization?
- ▽ Network Analysis: *With Whom*
  - Does the visualization contain information about social networks?

Audience

- ♂ Gender – are we targeting a certain gender?
- ⑤ Age – is it intended for certain age groups?
- ✍ Education – is the level of education important
- ♿ Disability – are disabilities taken into account (for example colour blindness)?
- ☐ Contextual parameters, e.g.
  - ☕ Leisure – related to our leisure
  - 🏭 Business – related to business
  - ☣ Scientific – related to science
  - † Religious – related to religion
  - ☐ Any other information defining the audience

Medium

- ✎ Printed medium
- ⌨ Digital medium
- ✇ Time-based – visualizing information using time
- ⊙ Location-based – spatially visualizing information
- Ⓐ Modality Text – contains text
- 🔔 Modality Sound – contains sound
- ☞ Interactive visualization
- ☐ Other – other information about the medium

Framework

Level

- ○ Micro level
- ◐ Meso level
- ● Macro level

Type

- 🛈 Profiling
- 🕒 Temporal
- 🌍 Geospatial
- ≡ Topical
- ▽ Network

Audience

- ♂ Gender
- ⑤ Age
- ✍ Education
- ♿ Disability
- ☐ Context, e.g.
  - ☕ Leisure
  - 🏭 Business
  - ☣ Scientific
  - † Religious
  - ☐ Other

Medium

- ✎ Printed
- ⌨ Digital
- ✇ Time-based
- ⊙ Spatial
- Ⓐ With Text
- 🔔 With Sound
- ☞ Interactive
- ☐ Other

4 Perception

4.1 Physiology

Gesichtsfeld

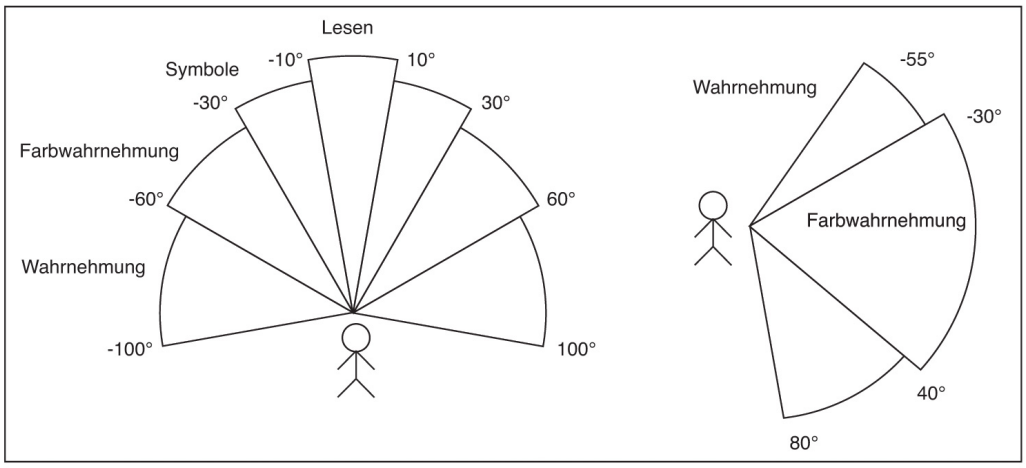
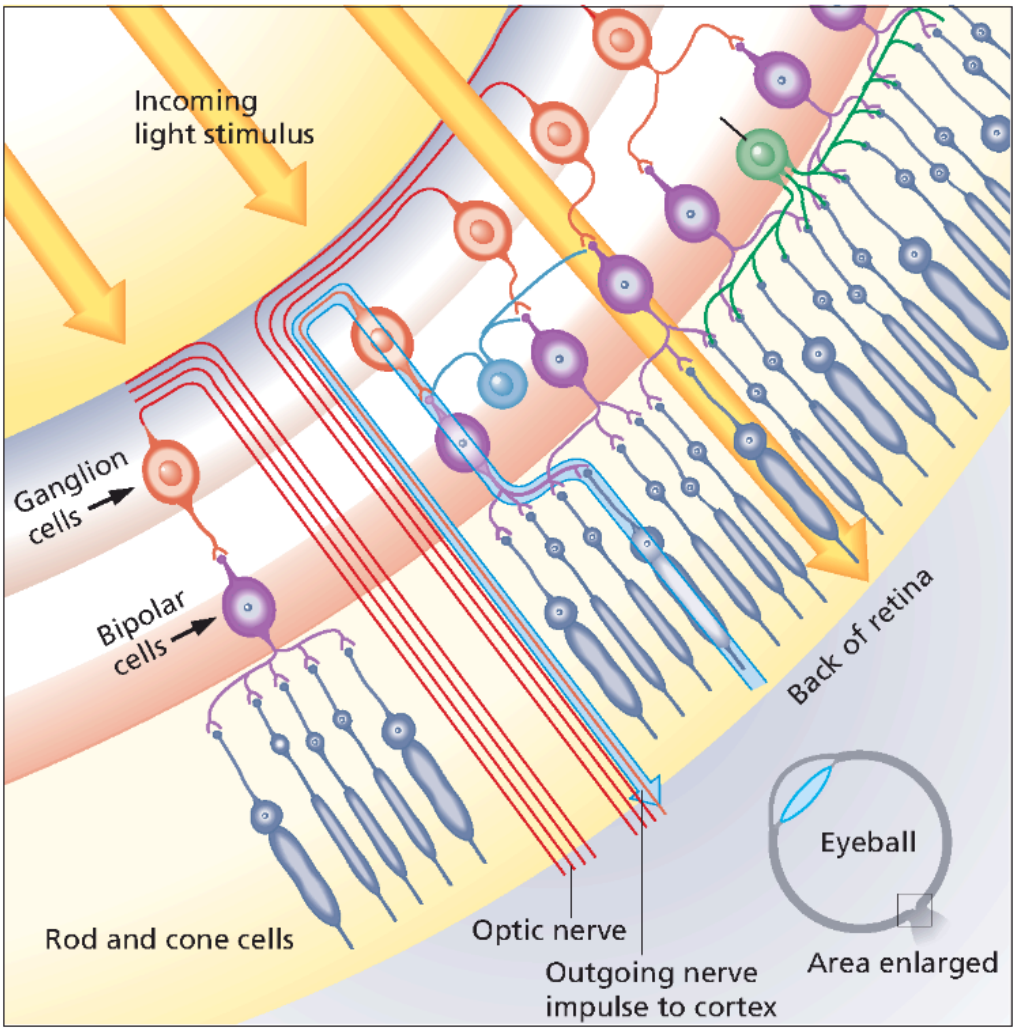


Abbildung 1.1: Sehfeld beim Menschen. Links horizontal, rechts vertikal (nach Herczeg, 1994)

Source: Malaka et al. (2009)

- Höchste Auflösung in der Fovea in der Mitte des Sehfeldes
- Dort finden sich viele Zapfen, aber keine Stäbchen
- Nachts sind wir im Zentrum des Sehfeldes faktisch blind
- In der Peripherie ist das Sehen stark eingeschränkt

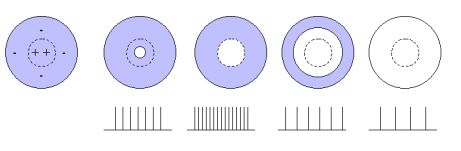
Retina



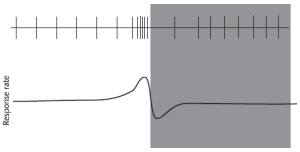
Source: Zimbardo et al. (2012)

Rezeptives Feld

- Bei Ganglionzellen ist das rezeptive Feld rund
- Das rezeptive Feld wird in ein Zentrum und ein Umfeld unterteilt und man unterscheidet On-Zentrum-Neurone und Off-Zentrum-Neurone
  - On-Zentrum-Neuronen haben ein erregendes Zentrum und ein hemmendes Umfeld
  - Bei Off-Zentrum-Neuronen verhält es sich umgekehrt
- Durch Erregung und Hemmung wird die Feuerrate des Neurons manipuliert



Licht (weiß), On-Zentrum
Source: Jänicke (2016)



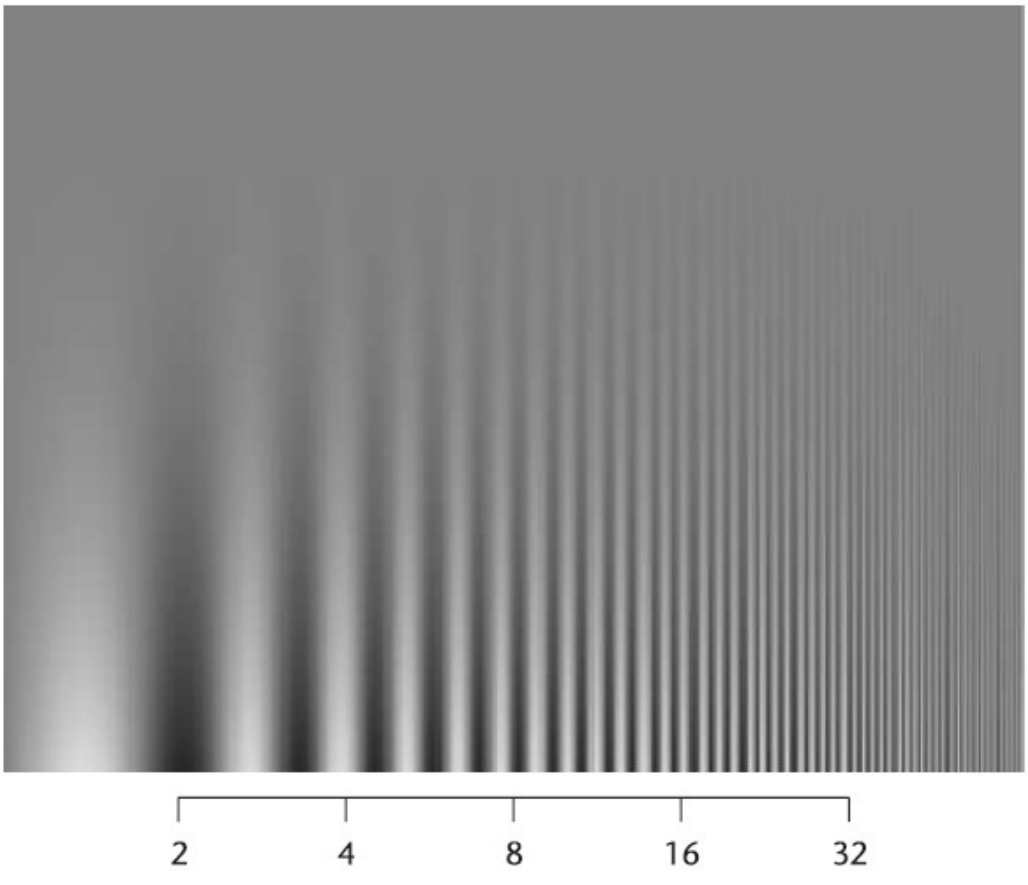
Kante, On-Zentrum
Source: Ware (2004)

Optische Täuschungen

Mit dieser Theorie kann man einige optische Täuschungen erklären

- Hermann Gitter (links): Schwarze Punkte erscheinen an den Schnitten weißer Geraden
- Kontrast Illusion (rechts): Abhängig von der Hintergrundfarbe wird ein und derselbe Grauton unterschiedlich wahrgenommen

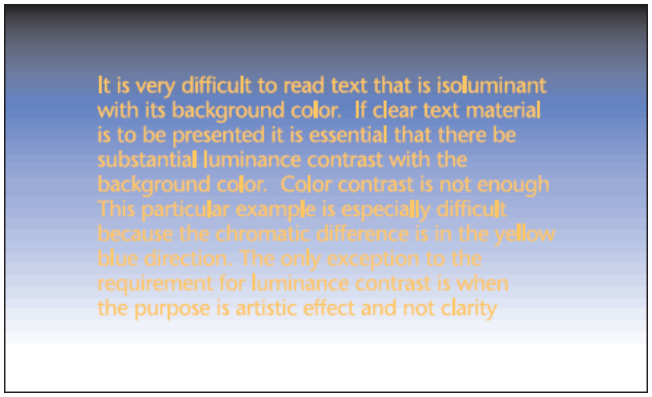
Spatial Contrast



Source: Ware (2004)

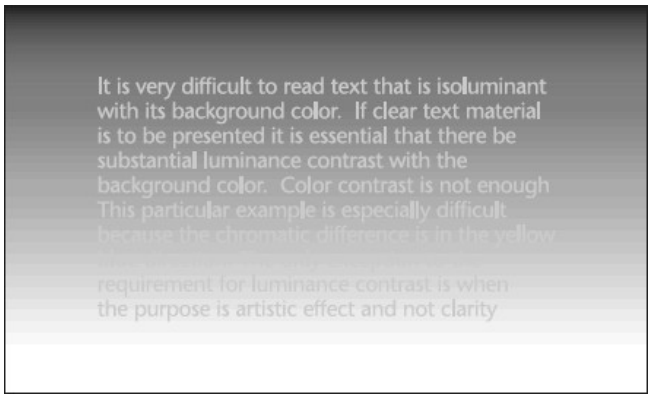
4.2 Color

Isoluminanz



Source: Ware (2004)

Isoluminanz



Source: Ware (2004)

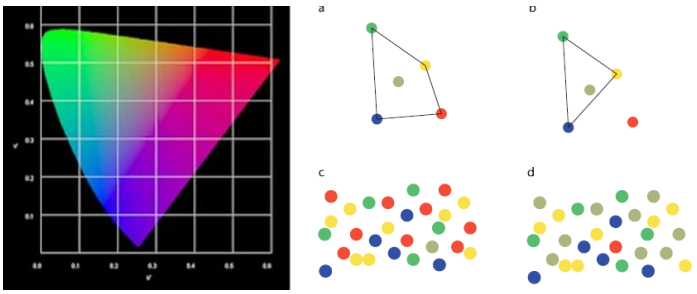
Farbkodierung von Objekten

Bei der Verwendung von Farbe zur Unterscheidung von Merkmalen müssen einige Punkte beachtet werden:

- Unterscheidbarkeit
- Eindeutige Farbtöne
- Kontrast zum Hintergrund
- Farbschwäche
- Anzahl
- Größe der Farbfläche
- Konventionen

Farbkodierung von Objekten

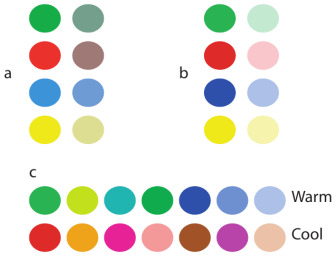
- **Unterscheidbarkeit**: Die Farben sollen leicht voneinander zu unterscheiden sein
  - Wenn es darum geht ein Objekt einer bestimmten Farbe schnell zu finden, sollte diese außerhalb der konvexen Hülle der anderen Farben liegen



Source: Jänicke (2016); Ware (2004)

Farbkodierung von Objekten

- **Eindeutige Farbtöne**: Gegenfarben haben in den meisten Kulturen und Sprachen einen eigenen spezifischen Namen und werden leicht erkannt
  - Zu bevorzugen, wenn nur wenige Farben benötigt werden
  - Wenn möglich nicht mehrere Farben aus der gleichen Farbfamilie verwenden
  - Gegenfarben: Blau-Gelb, Rot-Grün, Schwarz-Weiß



Families of colors: (a) Pairs related by hue, family members differ in saturation. (b) Pairs related by hue, family members differ in saturation and lightness. (c) A family of warm hues and a family of cool hues.

Source: Ware (2004)

Farbkodierung von Objekten

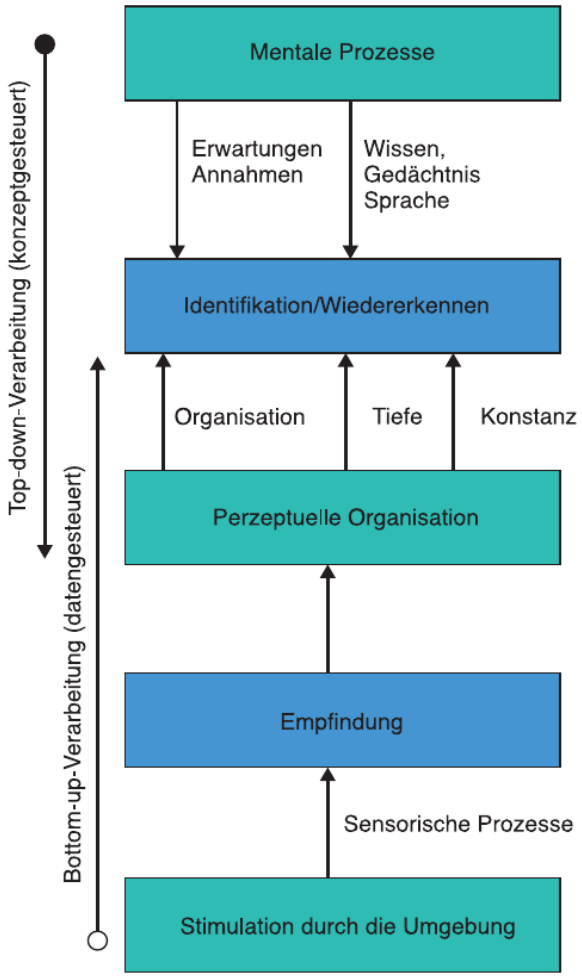
- **Kontrast zum Hintergrund**: Es muss beachtet werden, dass Farben auf unterschiedlichem Hintergrund unterschiedlich wirken können
  - Wechselwirkungen können durch eine einheitliche Kontur (z.B. schwarz oder weiß) verkleinert werden
  - Isoluminanz zwischen Objekt und Hintergrund ist zu vermeiden
- **Farbschwäche**: Da es relativ viele Menschen mit Farbschwäche gibt sollten Farbkodierung basierend auf rot-grün Kontrasten vermieden werden
- **Anzahl**: Nur 5 bis 10 Farben können schnell unterschieden werden

Farbkodierung von Objekten

- **Größe der Farbfläche**: Die Größe der farblich kodierten Objekte sollte nicht zu klein sein, da sie sonst nicht unterschieden werden können.
  - Allgemein gilt: Für kleine Farbflächen sollten stark gesättigte und stark unterschiedliche Farben verwendet werden, für große Flächen eher Farben mit niedrigerer Sättigung und geringerem Abstand
  - Bei farbig hinterlegtem Text sollte eine helle Farbe gewählt werden
- **Konventionen**: Einige Farben haben bestimmte Bedeutungen
  - Rot = heiß oder Gefahr – Blau = kalt – Grün = Leben
  - Man beachte: Andere Länder, andere Sitten!
    - z.B. in China gilt rot = Leben oder Glück und grün = Tod

4.3 Processing Pipeline

Verarbeitungsprozesse



- Verarbeitung visueller Information komplexer Prozess
- Man unterscheidet grob drei Stufen der Verarbeitung:
  1. Sensorische Prozesse
     - Parallele Erfassung grundlegender Merkmale
  2. Perzeptuelle Organisation
     - Mustererkennung
  3. Aufgabenorientierte Verarbeitung
     - Identifikation
     - Wiedererkennen

Source: Ware (2004), Gerrig and Zimbardo (2008, Graphik)

Stufe 1: Sensorische Prozesse

- Milliarden Neuronen erfassen gleichzeitig unterschiedliche Merkmale des visuellen Feldes, z.B. Helligkeit, Farbe und Orientierung von Kanten
- Diese initiale Verarbeitung ist zum größten Teil unabhängig davon, worauf wir unsere Aufmerksamkeit richten
- Wichtige Merkmale:
  - Schnelle parallele Verarbeitung
  - Extraktion fundamentaler Merkmale
  - Information wird nur kurz gespeichert
  - Datenbasiertes bottom-up Modell der Verarbeitung
- In der ersten Stufe kann sehr viel visuelle Information parallel verarbeitet werden
- Kann genutzt werden um die Aufmerksamkeit zu lenken; bestimmte Aspekte hervorzuheben
- So kann man den Betrachter dabei unterstützen wichtige Informationen schnell zu erkennen.

Stufe 2: Perzeptuelle Organisation

- Schätzungen der wahrscheinlichen Größe, Form, Bewegung, Entfernung und Ausrichtung eines Objekts
- Schätzungen basieren auf mentalen Berechnungen, die Vorwissen mit aktueller Evidenz aus den Sinnen sowie dem Reiz in seinem Wahrnehmungskontext kombinieren

- Synthese einfacher sensorischer Merkmale wie beispielsweise Farben, Kanten und Linien zu einem Perzept eines Objekts
- Wichtige Merkmale:
  - Langsame serielle Verarbeitung
  - Verwendung von Kurzzeit- und Langzeitgedächtnis
  - Wechsel zwischen Merkmalsverarbeitung (bottom-up) und Aufmerksamkeit (top-down)
  - Symbole erhalten komplexere Bedeutungen
  - Verschiedene Verarbeitungspfade
    * Objekterkennung – what-system
    * Bewegungssteuerung – action-system, where-system

Stufe 3: Aufgabenorientierte Verarbeitung

- Weist den Perzepten Bedeutung zu
- Runde Objekte ''werden'' zu Fußbällen, Münzen, Uhren, Orangen oder Monden; Menschen werden als weiblich oder männlich identifiziert, Freund/Feind, Verwandter/Star
- Die Aufmerksamkeit wird gezielt auf relevant Aspekte des visuellen Feldes gerichtet und wenige relevante Objekte werden im Kurzzeitgedächtnis gespeichert
- Wichtige Merkmale:
  - Langsame serielle Verarbeitung
  - Verwendung von Kurzzeit- und Langzeitgedächtnis
  - Top-down Verarbeitung
  - Verarbeitung richtet sich nach der Fragestellung
- Verschiedene Objekte in einer Visualisierung sollten deutlich unterscheidbar sein, um diesen Prozess zu beschleunigen (vergleiche ''Wo ist Walter?'')

4.4 Attention

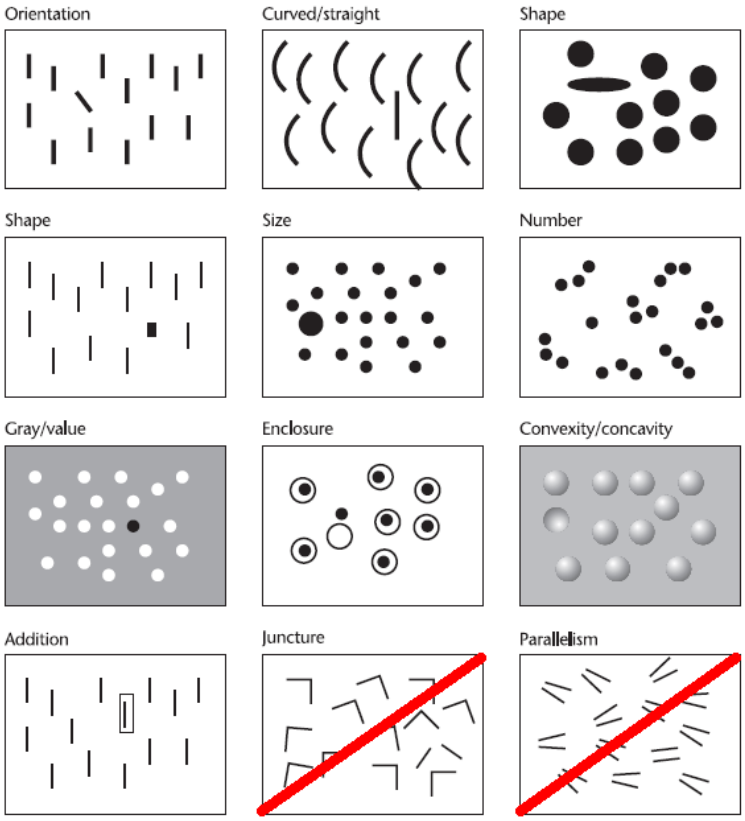
Präattentive Verarbeitung: Definition

> Präattentive Verarbeitung: Die Verarbeitung sensorischer Informationen, die einer Aufmerksamkeitszuwendung auf spezifische Objekte vorausgeht (Gerrig and Zimbardo, 2008).

- Die erste Stufe der Verarbeitung visueller Information erfasst das gesamte visuelle Feld
- Dieser Schritt wird präattentiv genannt, da in ihm Informationen erfasst werden noch bevor Aufmerksamkeit (attention) darauf gerichtet wird
- Ob ein Reiz präattentiv ist wird experimentell bestimmt, indem man die Zeit misst, die Testpersonen brauchen um den Zielreiz in einer Menge von Distraktoren zu finden

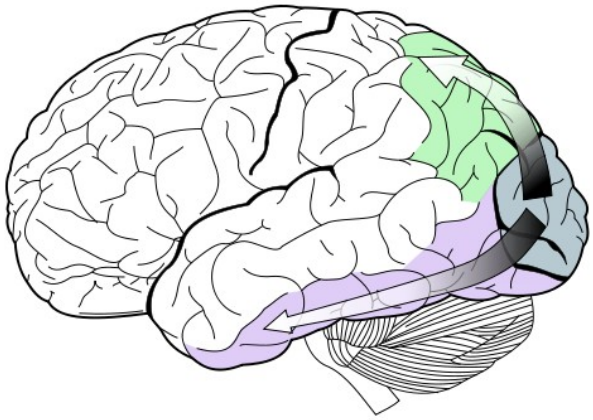
Präattentive Wahrnehmung: Beispiele

- Form:
  - Ausrichtung
  - Größe
  - Krümmung
  - Länge & Breite von Linien
  - Anzahl
  - Annotationen
- Farbe:
  - Farbton, Intensität
- Räumliche Position:
  - Konkav, Konvex
  - Einschluß
- Nicht Parallelität
- Nicht Verbindung



Source: Ware (2004)

Two-Streams-Theory



Source: Ware (2004)

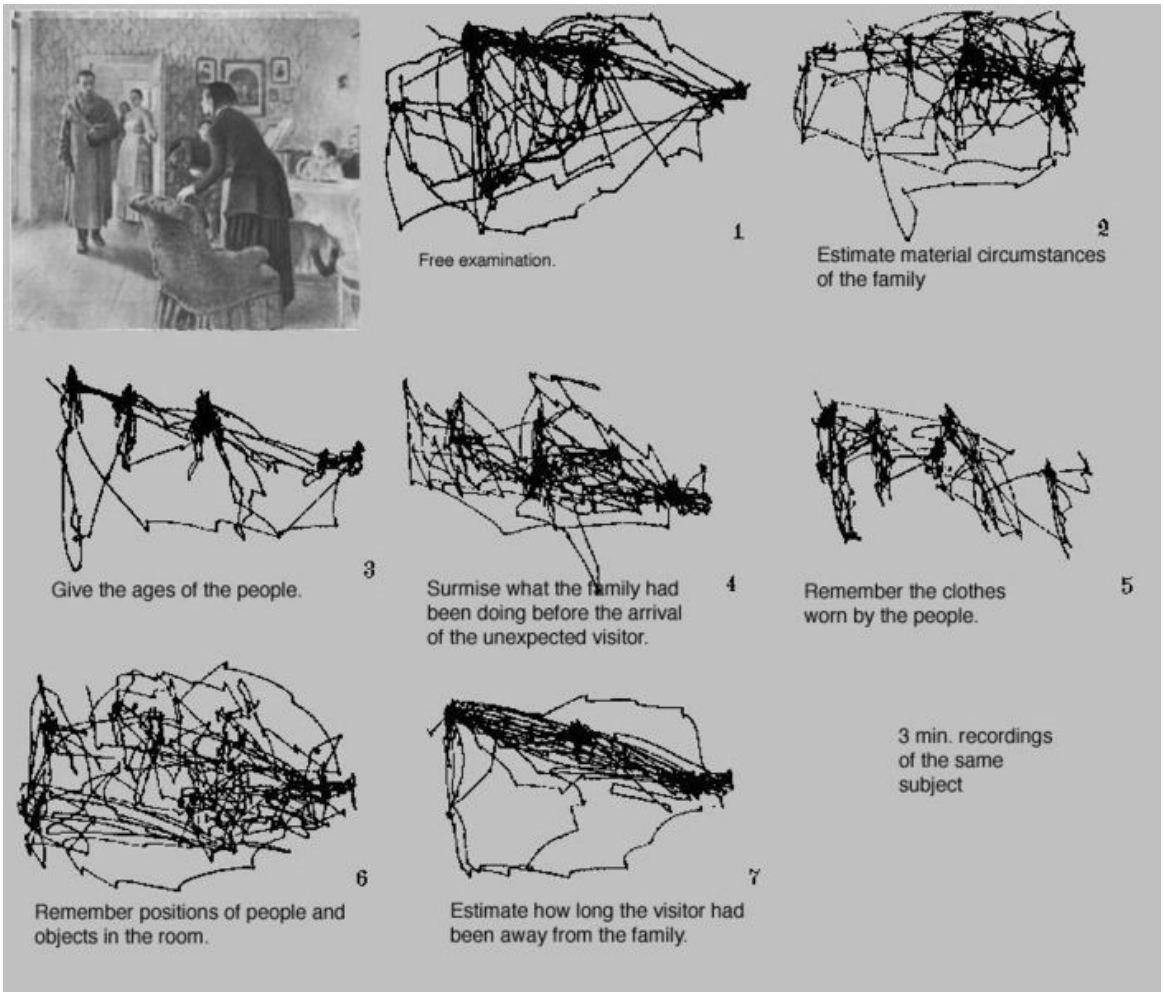
Nach der ''Two-Streams Theorie'' wird die optische Information danach in zwei Systemen weiterverarbeitet, dem dorsalen ''Wo-System'' und dem ventralen ''Was-System''

- Das dorsale System (grün) ist u.a. für die Wahrnehmung von Bewegung, Tiefe, räumlicher Organisation und für die Planung von Handlungen (z.B. Greifen) verantwortlich
- Das ventrale System (lila) gleicht die aufgenommene Information mit vorhandenem Wissen ab und ordnet das Gesehene ein – es ist u.a. verantwortlich für Wahrnehmung von Objekten, Formen und Gesichtern

Scheinwerfertheorie

- Die Scheinwerfertheorie erklärt wie die Aufmerksamkeit beim Betrachten einer Szene gesteuert wird
- Grundannahme ist dass die Aufmerksamkeit wie ein Scheinwerfer ist, der verschiedene Aspekte einer Szene beleuchten kann
- Fällt die Aufmerksamkeit des Betrachters auf einen kleinen Teil der Szene, kann man dort genaue Details wahrnehmen

- Die Verarbeitung erfolgt seriell, so dass der Aufmerksamkeitsscheinwerfer von einem Punkt zum nächsten geleitet wird
- Der "Weg" der Aufmerksamkeit durch die Szene ist kontextabhängig, z.B. ausgehend von der Aufgabe, die gelöst werden soll



Source: Jänicke (2016)

Gestaltgesetze (-prinzipien)

- Gesetz der Nähe
  - gruppiert Dinge zusammen, die räumlich oder zeitlich nah sind
- Gesetz der Ähnlichkeit/Gleichheit
  - gruppiert Bildteile, die nach Farbe, Form, Helligkeit, Größe, Orientierung ähnlich sind
- Gesetz der guten Fortsetzung
  - präferiert räumliche oder zeitliche Einfachheit
- Gesetz der Geschlossenheit
  - neigt dazu, kleine Lücken aufzufüllen
- Gesetz des gemeinsamen Schicksals
  - Objekte werden gerne als Gruppen wahrgenommen
- Weiterhin komplexere Prinzipien
  - Gesetz der Symmetrie
  - Unterscheidung von Vorder- und Hintergrund

5 Data

5.1 Communicate

Purpose

- Moving on: What is the purpose of your visualization?
- Be clear about the motivation behind a project's inception
- Involves identifying who it is for and what needs you are trying to fulfill
- What is the intention behind your project and how do you define the visualization's function and tone
- Identify and assess the impact of the additional key factors that will have an effect on your project
  - Helps you surface all the restrictions, characteristics, and requirements surrounding your project that will determine how you tackle it
- What is a purpose?
  - reason for existing
  - intended effect

Intent: The Visualization's Function

- The intended function of a data visualization concerns the functional experience you create between your design, the data, and the reader/user
- We can form three separate clusters or categories of function
- While there is always a chance of slight overlap, there will be a significant difference in your design choices depending on whether the function of your visualization is to:
  - Convey an *explanatory* portrayal of data to a reader
  - Provide an interface to data in order to facilitate visual *exploration*
  - Use data as an *exhibition* of self-expression

Explain

- Explanatory data visualization is about conveying information to a reader in a way that is based around a specific and focused narrative
- Editorial approach to synthesize the requirements of your target audience with the key insights and most important analytical dimensions you are wishing to convey
- Different approaches:
  - Information dashboard in a corporate setting (performance figures with problems highlighted)
  - A graphic in a newspaper, explaining the complexity and severity of the problems around the economic crisis
  - An animated design to display patterns of population migration over time
  - Physical or ambient visualization designed to draw attention to the sugar content of certain drinks
- The end result is typically a visual experience built around a carefully constructed narrative

Explore I

- Exploratory data visualization design is slightly different: we are seeking to facilitate the familiarization and reasoning of data through a range of user-driven experiences
- In contrast to explanatory-based functions, exploratory data visualizations lack a specific, single narrative
- They are more about visual analysis than just the visual presentation of data
- Exploratory solutions aim to create a tool, providing the user with an interface to visually explore the data
- They can seek out personal discoveries, patterns, and relationships, thereby triggering and iterating curiosities
  - Opens up the possibility for chance or serendipitous findings caused by forming different combinations of variable displays

Explore II

- The key feature that differentiates an exploratory piece from an explanatory piece is the amount of work you have to do as a reader to discover insights
  - For explanatory pieces, the designer should do the hard work and create a clear portrayal of the interesting stories and analysis from a dataset
  - An exploratory piece will be more about the readers doing the analysis themselves, putting the effort in to discover things that strike them as being significant or interesting

Exhibit

- Designs that use data as the raw material, but where the intention is somewhat removed from a pure desire to inform
- Rather, the objective is closer to a form of exhibition or self-expression through data representation
- This genre of work embodies the term ''data art''
- Characterized by a lack of structured narrative and absence of any visual analysis capability
- Instead, the motivation is much more about creating an artifact, an aesthetic representation or perhaps a technical/technique demonstration
- In the following example, we see an example of ''data art'' that visualizes all the adjectives used in Cormac McCarthy's book ''The Road''
  - Adjectives arranged radially in alphabetical order, each line represents a timeline of the book, beginning at perimeter

Science vs. Art

- ''Science''
  - Concerned with preserving the efficiency and accuracy of judgments derived from a visualization
  - Variations in data representation that steer away from this goal are believed to reduce the quality and effectiveness of a visualization
- ''Art''
  - Concerned with experimentation, finding creative expressions of data, and new aesthetic connections with an audience
- The latter enhances the field by demonstrating what can be achieved through the aesthetic and technological creativity
- The former help us understand what we should do through the pursuit of evidence and observation of rules around human cognition and visual perception

Pragmatic and Analytical

Jock Mackinlay
A visualization is more effective than another visualization if the information conveyed by one visualization is more readily perceived than the information in the other (in: Kirk (2012)).

- Designs that fit this classification will often involve data being represented through the use of bar charts, line charts and dot plots, for example
- Stylistically, they will be characterized by a rather clinical look-and-feel that is consistent with the next sample image, taken from a project analyzing Olympic results over the years

Emotive and abstract

Chris Jordan
I have a fear that we aren't feeling enough, we aren't able to digest these huge numbers (in: Kirk (2012)).

- Abstract visualization, in terms of its tone, is more about creating an aesthetic that portrays a general story or sense of pattern
- You might not be able to pick out every data point or category, but there is enough visual information to give you a feel for the physicality of the data
- This next image visualizes the global airline transportation network
- The project was designed to assess the threat of infectious diseases

Rehash

- Intent: The Visualization's Function
  - Explain
  - Explore
  - Exhibit
- Intent: The Visualization's Tone
  - Pragmatic and analytical
  - Emotive and abstract

5.2 Process

Process: Fry

- **Acquire:** Obtain the data, whether from a file on a disk or a source over a network.
- **Parse:** Provide some structure for the data's meaning, and order it into categories.
- **Filter:** Remove all but the data of interest.
- **Mine:** Apply methods from statistics or data mining as a way to discern patterns or place the data in mathematical context.
- **Represent:** Choose a basic visual model, such as a bar graph, list, or tree.
- **Refine:** Improve the basic representation to make it clearer and more visually engaging.
- **Interact:** Add methods for manipulating the data or controlling what features are visible.

Source: Fry (2008)

Process: Yau

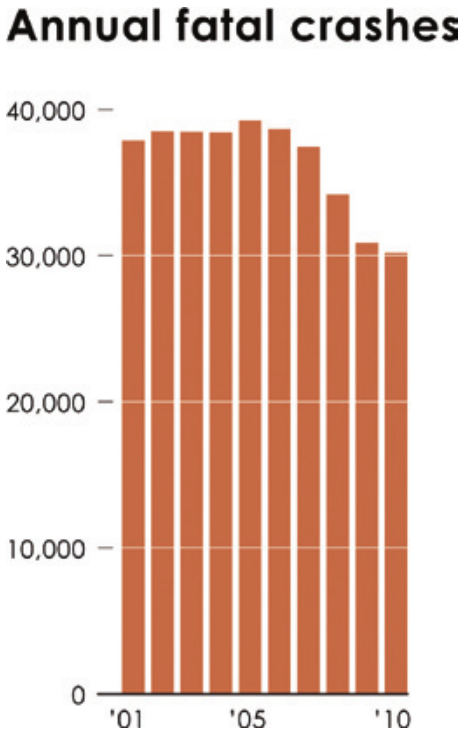
- What data do you have?
- What do you want to know about your data?
- What visualization methods should you use?
- What do you see and does it makes sense?

5.3 Aspects of Data

Rooted in Data

Rooted in Data
''Visualization is often thought of as an exercise in graphic design or a brute-force computer science problem, but the best work is always rooted in data. To visualize data, you must *understand what it is*, what it *represents in the real world*, and in *what context you should interpret it* in. Data comes in different shapes and sizes, at various granularities, and with uncertainty attached, which means totals, averages, and medians are only a small part of what a data point is about. It twists. It turns. It fluctuates. It can be personal, and even poetic. As a result, you can find visualization in many forms.'' (Yau, 2013)

Time-Mapping



Source: Yau (2013)

- A look at crashes over time shifts focus to the events themselves
- This Figure shows the number of accidents per year, which tells a different story than the total seen before
- Accidents still occurred in the tens of thousands annually, but there was a significant decline from 2006 through 2010, and fatalities per 100 million vehicle miles traveled (not shown) also decreased

Granularity

- Seasonal cycles become obvious at month-by-month granularity, as shown in the next figure
- Incidents peak during the summer months when people go on vacation and spend more time outside, whereas
- during the winter, fewer people drive, so there are fewer crashes
- This happens every year
- At the same time, you can still see the annual decline overall between 2006 and 2010.

Zooming In

- You can increase granularity to crashes by the hour
- The next figure breaks it down
- Each row represents a year, so each cell in the grid shows an hourly time series for the corresponding month
- With the exception of a new year's spike during the midnight hour, it's hard to make out patterns at this level because of the variability
  - Actually, the monthly chart is hard to interpret, too, if you don't know what you're looking for

Aggregation

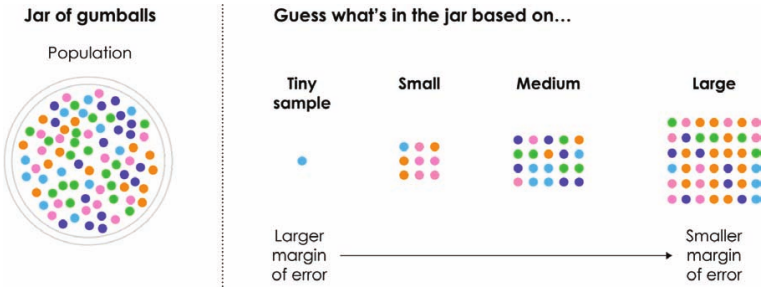
- There are clear patterns, though, if you aggregate, as shown in the next figures
- Instead of showing values at every hour, day, or month, you can aggregate on specific time segments to explore the distributions
- What was hard to discern, or looked like noise before, is easy to see here

Uncertainty

- There are a lot of examples for data with uncertainty
  - Weather reports
  - Time to complete a file transfer
  - Remaining battery time
- When you have data that is a series of means and medians or a collection of estimates based on a sample population, you should always wonder about the uncertainty

Example 1. The United States Census Bureau releases data about the country on topics such as migration, poverty, and housing, which are estimates based on samples from the population. A margin of error is provided with each estimate, which means that the actual count or percentage is likely within a given range

Sample Size



Source: Yau (2013)

- Uncertainty in statistical data can be reduced by using an appropriate sample size
- The needed sample size for a target uncertainty can be computed

Context makes Data Useful

- Without context, data is useless, and any visualization you create with it will also be useless
- Using data without knowing anything about it, other than the values themselves, is like hearing an abridged quote secondhand and then citing it as a main discussion point in an essay
  - It might be okay, but you risk finding out later that the speaker meant the opposite of what you thought
- You have to know the metadata, or the data about the data, before you can know what the numbers are actually about

Questions

- Who
  - collects the data
  - is the data about
- How was it collected
- What was collected
- When was it collected
- Where was it collected
- Why was the data collected

Ethical Questions

- Is it always OK to make a visualization?
- Consider the following case
  - In 2010, Gawker Media, which runs large blogs like Lifehacker and Gizmodo, was cracked, and 1.3 million usernames and passwords were leaked
  - They were downloadable via BitTorrent
  - The passwords were encrypted, but the attackers cracked about 188,000 of them, which exposed more than 91,000 unique passwords
  - What would you do with that kind of data?

What before How

- Next question: what is it we are trying to say with the visualization we are developing?
- We first need to determine what are the specific messages we are looking to communicate to our audience – the *what*
- The *how* this is said will be covered in the design stage
  - This is roughly equivalent to a user-centred design process: before we look at how the application looks like, we first need to understand what the application should offer to the user
- **Editorial focus:** An editorial approach to visualization design requires us to take responsibility to filter out the noise from the signals and to identify the most valuable, most striking, or most relevant dimensions of the subject matter

5.4 Data Preparation

Steps

- Acquisition
- Examination
  - Completeness
  - Quality
- Data Types
- Transformation
  - For Quality
  - For Analysis
- Consolidation

Acquisition

- First, you need to get hold of your data
- As discussed, this might already be provided to you from those commissioning the work
- You might have independently formed a sense of the specific subject dimensions on which you require data
- Alternatively, it may be that you have yet to focus beyond a broad subject level
  - Obtained from a colleague, client, or other third-party entity

- A download taken from an organizational system
- Manually gathered and recorded
- Extracted from a web-based API
- Scraped from a website
- Extracted from a Documents (such as PDF files)

Examination

- Once we've got the data, a thorough examination will determine your level of confidence in the suitability of what you have acquired
- This involves assessing the completeness and fitness of the data to potentially serve your needs
- will enable you to quickly scan, filter, sort, and search through your data set
- Potential issues:
  - Completeness
  - Quality

Examination: Completeness

- Is it all there or do you need more?
- Is the size and shape consistent with your expectations?
- Does it have all the categories you were expecting?
- Does it cover the time period you wanted?
- Are all the fields or variables included?
- Does it contain the expected number of records?

Examination: Quality

- Are there noticeable errors?
- Are there any unexplained classifications or coding?
- Any formatting issues such as unusual dates, ASCII characters?
- Are there any incomplete or missing items?
- Any duplicates? Does the accuracy of the data appear fine?
- Are there any unusual values or obvious outliers?

Data types: Categories

Categorical nominal	Countries, gender, text
Categorical ordinal	Olympic medals, "Likert" scale
Quantitative (interval-scale)	Dates, temperature
Quantitative (ratio-scale)	Prices, age, distance

Data types: Operations

N – Nominal (labels) — Fruits: Apples, oranges, . . .
Operations: $= \neq$

O – Ordered — ECTS Grades A, B, C, . . .
Operations: $= \neq < > \leq \geq$

Q – Interval (location of 0 arbitrary) — Dates: 19. Jan 2017
Loc.: (LAT 33.98, LON -118.45)
Operations: $= \neq < > \leq \geq -$
Like a geometric point. Cannot compare directly. Only differences (i.e. intervals) may be compared.

Q – Ratio (location of 0 fixed) — Measurements: Length, Temp, . . .
Counts and amounts
Operations: $= \neq < > \leq \geq - \div$
Like a geometric vector, origin is meaningful.

Transformation for Quality

- This task is about tidying and cleaning your data in response to the examination stage above
- We are looking to resolve any of the errors we discovered in order to transform the condition of the data we're going to be working with for our design
- Plugging the gaps caused by missing data, removing duplicates, cleaning up erroneous values, and handling uncommon characters are some of the treatments we may be required to apply

Transformation for Analysis

- Here, we focus on preparing and refining it in anticipation of its intended use for analysis and presentation
  - Parsing (split up) any variables, such as extracting year from a date value
  - Merging variables to form new ones, such as creating a whole name out of title, forename, and surname
  - Converting qualitative data/free-text into coded values or keywords
  - Deriving new values out of others, such as gender from title or a sentiment out of some qualitative data
  - Creating calculations for use in analysis, such as percentage proportions
  - Removing redundant data for which you have no planned use (be careful)

Resolution: Choice

- **Full :** Plotting all data available as individual data marks
- **Filtered:** Exclude records based on a certain criteria
- **Aggregate:** ''Roll-up'' the data by, for instance, month, year, or specific category
- **Sample:** Apply (mathematical) selection rules to extract a fraction of your potential data
  - Particularly useful during a design stage if you have very large amounts of data and want to quickly develop mock-ups or test out ideas
- **Headline:** Just showing the overall statistical totals

Consolidation

- When you originally access your data, you will likely believe, or hope that you have everything you need
- However, it may be that after the examination and preparation work, you identify certain gaps in your subject matter
- Additional layers of data may be required to be combined (''mashed-up'') with our existing dataset, applied to perform additional calculations, or just to sit alongside this initial resource to help contextualize and enhance the scope of our communication
- Always spend a bit of time considering if there is anything else you anticipate needing to supplement your data to help frame the subject or tell the stories you want to communicate

5.5 Focus

Visual Design Options

- The way that you choose to represent your data – your selection of chart type – should be influenced by the questions you are trying to answer
- If you are asking a chart to facilitate a comparison between the values of different categories, you might deploy a bar chart
- You wouldn't use a line chart, unless you wanted to show how a value or values change over time
- A scatter plot can be the perfect method of comparing two quantitative values for different countries

Data Sketching

- Consider the potential of visualization for ourselves
- Visually analyzing a data set, and employing both inductive and deductive reasoning, enables us to learn more about our subject by exploring a dataset from all directions
- Rather than just looking at data, we are using visualization to actually see it, to find previously undiscoverable properties of our raw material, to learn about its shape, and the relationships that exists within
  - data sketching or pre-production visualization
- Using visualization techniques to become more intimate with our raw material and to start to form an understanding of what we might portray to others and how we might accomplish that

Exploration Dimensions

- Comparisons and proportions
  - E.g. using bar charts
- Trends and patterns
  - E.g. using line charts
- Relationships and connections
  - E.g. using scatter plots
- The chart types shown illustrative just a small section of the gallery of options we have to call upon

Deductive Approach

- Deductive reasoning involves confirming or finding evidence to support specific ideas
- A deductive approach to defining your data questions will involve a certain predetermined sense of what stories might be interesting, relevant, and potentially available within your data
- You are pursuing a curiosity by interrogating your data set in order to substantiate your ideas of what may be the key story dimensions

Inductive Approach

- Inductive reasoning is much more open-ended and exploratory
- We are not sure what the interesting stories might be
- We use analytical and visualization techniques to try and unearth potentially interesting discoveries, forming different and evolving combinations of data questions
- We may end up with nothing, we may find plenty
- Fundamentally, this is about using visual analysis to find stories

Comparisons and Proportions

- **Range and distribution:** Discovering the range of values and the shape of their distribution within each variable and across combinations of variables
- **Ranking:** Learning about the order of data in terms of general magnitude, identifying the big, medium, and small values.
- **Measurements:** Looking beyond just the order of magnitude to learn about the significance of absolute values
- **Context:** Judging values against the context of averages, standard deviations, targets, and forecasts

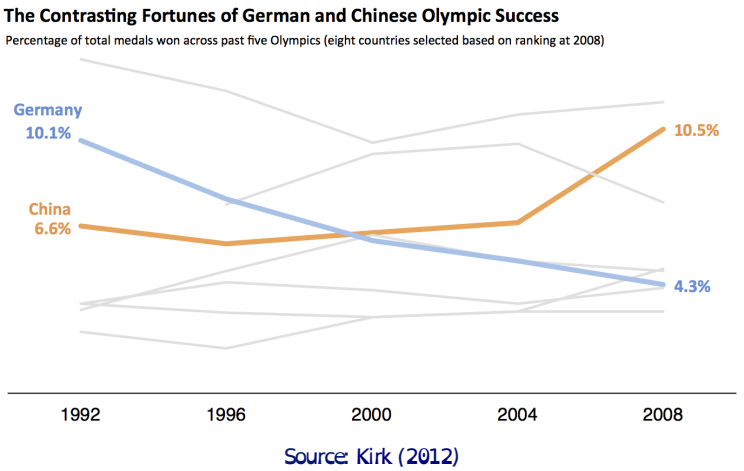
Trends and Patterns

- **Direction:** Are values changing in an upward, downward, or flat motion?
- **Rate of change:** How steep or flat do pattern changes occur? Do we see a consistent, linear pattern, or is it much more exponential in shape?
- **Fluctuation:** Do we see evidence of consistent patterns or is there significant fluctuation? Maybe there is a certain rhythm, such as seasonality, or perhaps patterns are more random
- **Significance:** Can we determine if the patterns we see are meaningful signals or simply represent the noise within the data?
- **Intersections:** Do we observe any important intersections or overlaps between variables, crossover points that indicate a significant change in relationship?

Relationships and Connections

- **Exceptions:** Can we identify any significant values that sit outside of the norm, such as outliers that change the dynamics of a given variable's range?
- **Correlations:** Is there evidence of strong or weak correlations between variable combinations?
- **Associations:** Can we identify any important connections between different combinations of variables or values?
- **Clusters and gaps:** Where is there evidence of data being "bunched"? Where are there gaps in values and data points?
- **Hierarchical relationships:** Determining the composition, distribution, and relevance of the data's categories and subcategories.

Example: Olympic Medals



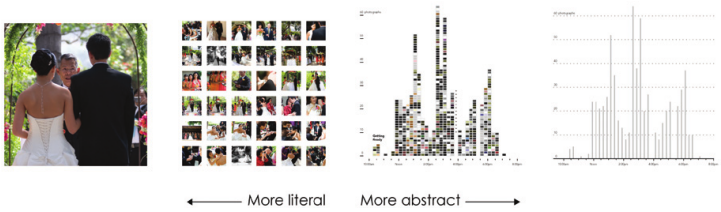
Source: Kirk (2012)

6 Representation

6.1 Components

Literal vs. Abstract

- When you visualize data, you represent it with a combination of visual cues that are scaled, colored, and positioned according to values
- Dark-colored shapes mean something different from light-colored shapes, or dots in the top right mean something different than dots in the bottom left
- Visualization is what happens when you make the jump from raw data to bar graphs, line charts, and dot plots
- Process taking you from a grid of photos to a bar graph

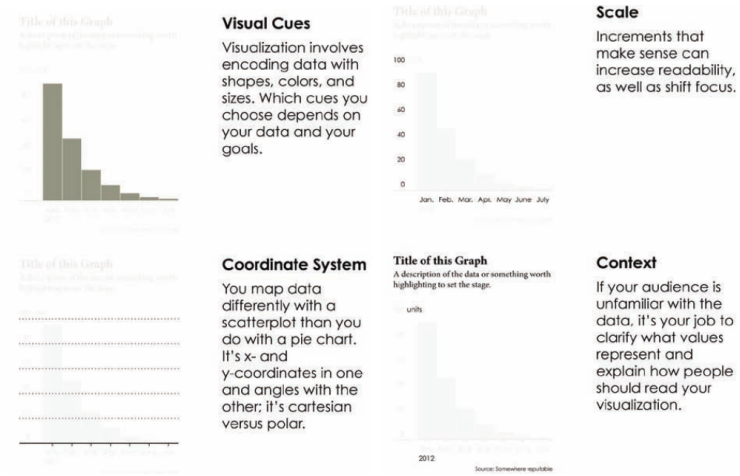


Source: Yau (2013)

Ingredients

- What are the ingredients of visualization?
- Breakdown into four components, with data as the driving force behind them:
  - visual cues,
  - coordinate system,
  - scale, and
  - context
- Each visualization, regardless of where it is on the spectrum, is built on data and these four components
- Sometimes, they are explicitly displayed, and other times they form an invisible framework
- The components work together, and your choice with one affects the others.

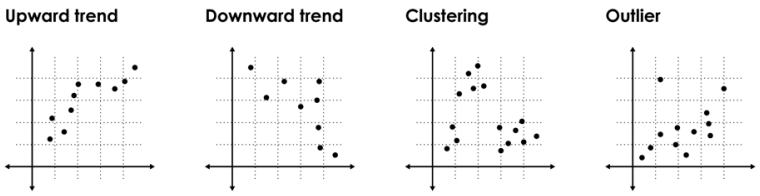
Ingredients



Source: Yau (2013)

Position

- When you use position as a visual cue, you compare values based on where others are placed in a given space or coordinate system
- For example, when you look at a scatterplot, as shown in the next Figure, you judge a data point based on its x- and y-coordinate and where it is relative to others



Source: Yau (2013)

Length

- Length is most commonly used in the context of bar charts
- The longer a bar is, the greater the absolute value, and it can work in all directions: horizontal, vertical, or even at different angles on a circle
- How do you judge length visually?
- You figure out the distance from one end of a shape to the other end, so to compare values based on length, you must see both ends of the lines or bars
- Otherwise, you end up with a skewed view of maximums, minimums, and everything in between
- As a simple example, as shown next, a major news outlet displayed a bar graph on television that compared a tax rate before and after a date

Angle

- For each angle in between zero and 360 degrees, there is an implied opposite angle that completes the rotation, and together those two angles are considered conjugates
- This is why angles are commonly used to represent parts of a whole, using the fan favorite, but often maligned, pie chart
- The sum of the wedges makes a complete circle
- Although the donut chart is often considered the pie chart's close cousin, arc length is the former's visual cue because the center of the circle, which indicates angles, is removed

Direction

- Direction is similar to angle, but instead of relying on two vectors joined at a point, direction relies on a single vector's orientation in a coordinate system
- You can see which way is up, down, left, and right and everything in between
- This helps you determine slope
- You can see increases, decreases, and fluctuations
- A rule of thumb is to scale your visualization so that direction fluctuates mostly around 45 degrees, but this is hardly a concrete rule
- The best thing to do is to start with this suggestion and then adjust accordingly based on context
- If a small change is significant, then it might be appropriate to stretch the scale so that you can see the shift
- In contrast, if a small change is not significant, don't stretch out the scale just to make a shift look dramatic.

Shapes

- Shapes and symbols are commonly used with maps to differentiate categories and objects
- Location on a map can be directly translated to the real world, so it makes sense to use icons to represent things in the real world
- You might represent forests with trees or residential areas with houses
- In a chart context, shapes to show variation are used less frequently than they used to be
- For example, triangles and squares could be used in a scatterplot, which is quicker to draw than to switch between colored pencils and pens or fill a single shape with a solid or cross-hatched pattern
- Nevertheless, varied shapes can provide context that points alone can't, and it's typically not more difficult to try with your favorite software

Area and Volume

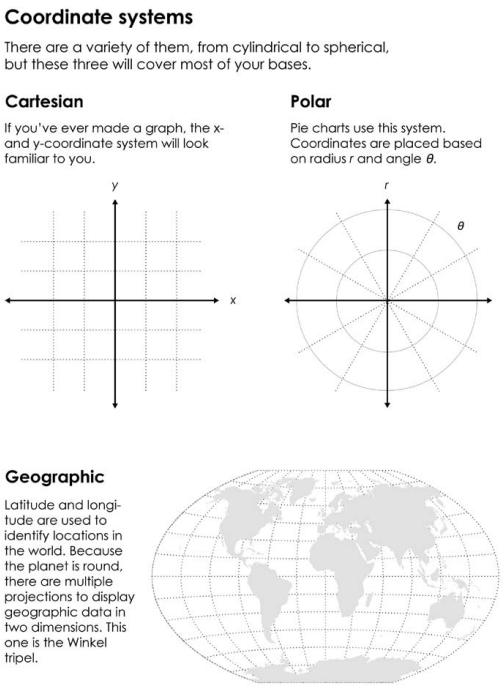
- Bigger objects represent greater values
- Like length, area and volume can be used to represent data with size, but with two and three dimensions, respectively
- For the former, circles and rectangles are commonly used, and with the latter, cubes and sometimes spheres
- You can also size more detailed icons and illustrations
- Be sure to mind how many dimensions you use
- The most common mistake is to size a two- or three-dimensional object by only one dimension, such as height, but to maintain the proportions of all dimensions
- This results in shapes that are too big and too small, which makes it impossible to fairly compare values

Color

- Color as a visual cue can be spilt into two categories: hue and saturation
- They can be used individually or in combination
- Color hue is what you usually just refer to as color
  - That's red, green, blue, and so on
- Differing colors used together usually indicates categorical data, where each color represents a group
- Saturation is the amount of hue in a color, so if your selected color is red, high saturation would be very red, and as you decrease saturation, it looks more faded
- Used together, you can have multiple hues that represent categories, but each category can have varying scales

6.2 Placement

Placement of Objects



Source: Yau (2013)

- When you encode data, you eventually must place the objects somewhere
- There's a structured space and rules that dictate where the shapes and colors go
- This is the coordinate system, which gives meaning to an x-y coordinate or a latitude and longitude pair
- There are several systems, but there are three that cover most of your bases: Cartesian, polar, and geographic

Cartesian

- The Cartesian coordinate system is the most commonly used one with charts
- You typically think of coordinates in the system as an x and y pair that is denoted as (x, y)
- Two lines that are perpendicular to each other, and range from negative to positive, form the axes
- The place the lines intersect is the origin, and the coordinate values indicate the distance from that origin
- You can also extend the Cartesian space to more than two dimensions
- The takeaway is that you can describe geometric shapes using Cartesian coordinates, which makes it easier to draw in the space
- From an implementation standpoint, the coordinate system enables you to encode values to paper or a computer screen

Polar

- The polar coordinate system consists of a circular grid, where the rightmost point is zero degrees
- The greater the angle is, the more you rotate counter-clockwise
- The farther away from the circle you are, the greater the radius is
- Place yourself on the outer-most circle, and increase the angle
- This rotates you counterclockwise toward the vertical line (or the y-axis if this were Cartesian coordinates), which is 90 degrees (that is, a right angle)
- Rotate one-quarter more, and you get to 180 degrees
- Rotate back to where you started, and that's a 360-degree rotation

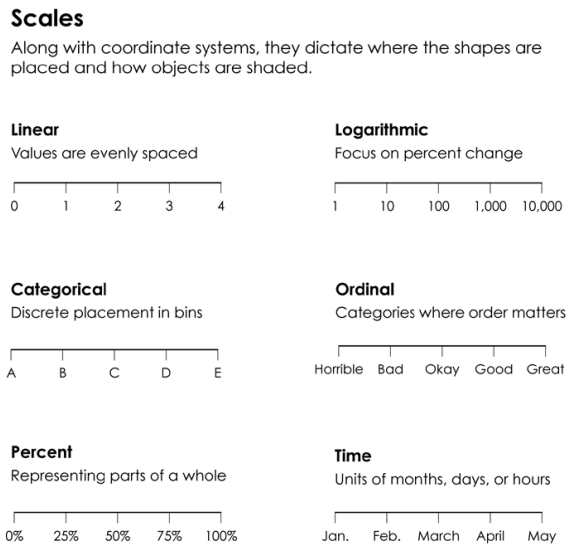
Geographic

- Location data has the added benefit of a connection to the physical world, which in turn lends instant context and a relationship to that point, relative to where you are
- A geographic coordinate system can map these points
- Location data comes in many forms, but it's most commonly described as latitude and longitude, which are angles relative to the Equator and Prime Meridian, respectively
  - Sometimes elevation is also included
- Latitude lines run east and west, which indicates north and south position on a globe
- Longitude lines run north and south and indicate the east and west position
- Elevation can be thought of as a third dimension
- Compared with Cartesian coordinates, latitude is like the horizontal axis, and longitude is like the vertical axis
  - That is, if you use a flat projection

Mapping Data

- Whereas coordinate systems dictate the dimensions of a visualization, scale dictates where in those dimensions your data maps to
- There's a variety of them, and you can even define your own scales based on mathematical functions, but most likely you'll rarely stray from the ones in the following Figure
- These can be grouped into three categories: quantitative/numerical, categorical, and time
  - Compare slide set "Data"

Scales: Example



Source: Yau (2013)

Quantitative

- The visual spacing on a linear scale is the same regardless of where you are on the axis
- So if you were to measure the distance between two points on the lower end of the scale, it'd be the same if they were at the high end of the scale
- On the other hand, a logarithmic scale condenses as you increase values

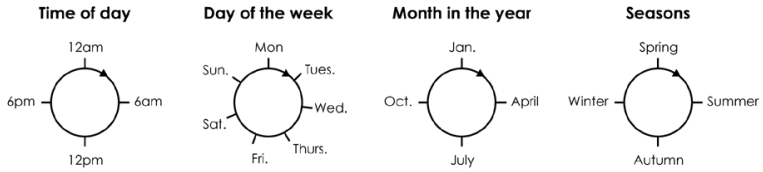
- This scale is used less than the linear scale and is not as well understood or straightforward for those who don't regularly work with data, but it's useful if you're interested in percent differences more than you are raw counts or your data has a wide range
- For example, when you compare state populations in the United States, you deal with numbers from the hundreds of thousands up to the tens of millions

Scale: Percentage

- A percent scale is usually linear, but when it's used to represent parts of a whole, its maximum is 100 percent
- As shown in the next Figure, the sum of all the parts is 100 percent
- This seems obvious – that the sum of percentages in a pie chart, represented with wedges, should not exceed 100 percent – but the mistake seems to come up occasionally
- Sometimes it's due to mislabeling, but some people just aren't familiar with the concept

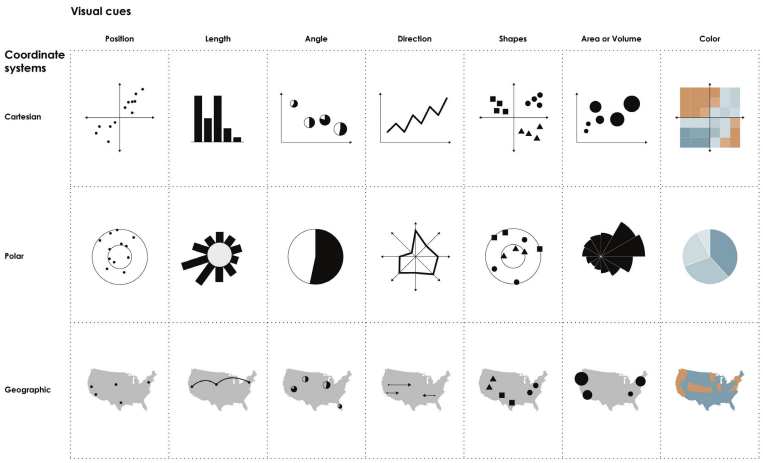
Time

- Time is a continuous variable, which lets you plot temporal data on a linear scale, but you can divide it into categories such as months or days of the week, which lets you visualize it as a discrete variable
- Also, it cycles
- There's always another noon time, Saturday, and January



Source: Yau (2013)

Visual Cues



Source: Yau (2013)

Walkthrough: Educational Attainment

Table 233. Educational Attainment by State: 1990 to 2009

[In percent. 1990 and 2000 as of April; 2009 represents annual averages for calendar year. For persons 25 years old and over. Based on the 1990 and 2000 Census of Population and the 2009 American Community Survey, which includes the household population and the population living in institutions, college dormitories, and other group quarters. See text, Section 1 and Appendix III. For margin of error data, see source]

State	1990			2000			2009		
	High school graduate or more	Bachelor's degree or more	Advanced degree or more	High school graduate or more	Bachelor's degree or more	Advanced degree or more	High school graduate or more	Bachelor's degree or more	Advanced degree or more
United States	**75.2**	**20.3**	**7.2**	**80.4**	**24.4**	**8.9**	**85.3**	**27.9**	**10.3**
Alabama	66.9	15.7	5.5	75.3	19.0	6.9	82.1	22.0	7.7
Alaska	86.6	23.0	8.0	88.3	24.7	8.6	91.4	26.6	9.0
Arizona	78.7	20.3	7.0	81.0	23.5	8.4	84.2	25.6	9.3
Arkansas	66.3	13.3	4.5	75.3	16.7	5.7	82.4	18.9	6.1
California	76.2	23.4	8.1	76.8	26.6	9.5	80.6	29.9	10.7
Colorado	84.4	27.0	9.0	86.9	32.7	11.1	89.3	35.9	12.7
Connecticut	79.2	27.2	11.0	84.0	31.4	13.3	88.6	35.6	15.5
Delaware	77.5	21.4	7.7	82.6	25.0	9.4	87.4	28.7	11.4
District of Columbia	73.1	33.3	17.2	77.8	39.1	21.0	87.1	48.5	28.0
Florida	74.4	18.3	6.3	79.9	22.3	8.1	85.3	25.3	9.0
Georgia	70.9	19.3	6.4	78.6	24.3	8.3	83.9	27.5	9.9
Hawaii	80.1	22.9	7.1	84.6	26.2	8.4	90.4	29.6	9.9
Idaho	79.7	17.7	5.3	84.7	21.7	6.8	88.4	23.9	7.5
Illinois	76.2	21.0	7.5	81.4	26.1	9.5	86.4	30.6	11.7
Indiana	75.6	15.6	6.4	82.1	19.4	7.2	86.6	22.5	8.1
Iowa	80.1	16.9	5.2	86.1	21.2	6.5	90.5	25.1	7.4
Kansas	81.3	21.1	7.0	86.0	25.8	8.7	89.7	29.5	10.2
Kentucky	64.6	13.6	5.5	74.1	17.1	6.9	81.7	21.0	8.5
Louisiana	68.3	16.1	5.6	74.8	18.7	6.5	82.2	21.4	6.9
Maine	78.8	18.8	6.1	85.4	22.9	7.9	90.2	26.9	9.6
Maryland	78.4	26.5	10.9	83.8	31.4	13.4	88.2	35.7	16.0
Massachusetts	80.0	27.2	10.6	84.8	33.2	13.7	89.0	38.2	16.4
Michigan	76.8	17.4	6.4	83.4	21.8	8.1	87.9	24.6	9.4
Minnesota	82.4	21.8	6.3	87.9	27.4	8.3	91.5	31.5	10.3
Mississippi	64.3	14.7	5.1	72.9	16.9	5.8	80.4	19.6	7.1
Missouri	73.9	17.8	6.1	81.3	21.6	7.6	86.8	25.2	9.5
Montana	81.0	19.8	5.7	87.2	24.4	7.3	90.8	27.4	8.3
Nebraska	81.8	18.9	5.9	86.6	23.7	7.3	89.8	27.4	8.8
Nevada	78.8	15.3	5.2	80.7	18.2	6.1	83.9	21.8	7.6
New Hampshire	82.2	24.4	7.9	87.4	28.7	10.0	91.3	32.0	11.2
New Jersey	76.7	24.9	8.8	82.1	29.8	11.0	87.4	34.5	12.9
New Mexico	75.1	20.4	8.3	78.9	23.5	9.8	82.8	25.3	10.4
New York	74.8	23.1	8.9	79.1	27.4	11.8	84.7	32.4	14.0
North Carolina	70.0	17.4	5.4	78.1	22.5	7.2	84.3	26.6	8.8
North Dakota	76.7	18.1	4.5	83.8	22.0	5.5	90.1	25.8	6.7
Ohio	75.7	17.0	5.9	83.0	21.1	7.4	87.6	24.1	8.8
Oklahoma	74.6	17.8	6.0	80.6	20.3	6.8	85.6	22.7	7.4
Oregon	81.5	20.6	7.0	85.1	25.1	8.7	89.1	29.2	10.4
Pennsylvania	74.7	17.9	6.6	81.9	22.4	8.4	87.9	26.4	10.2
Rhode Island	72.0	21.3	7.8	78.0	25.6	9.7	84.7	30.5	11.7
South Carolina	68.3	16.8	5.4	76.3	20.4	6.9	83.6	24.3	8.4
South Dakota	77.1	17.2	4.9	84.6	21.5	6.0	89.9	25.1	7.3
Tennessee	67.1	16.0	5.4	75.9	19.6	6.8	83.1	23.0	7.9
Texas	72.1	20.3	6.5	75.7	23.2	7.8	79.9	25.5	8.5
Utah	85.1	22.3	6.8	87.7	26.1	8.3	90.4	28.5	9.1
Vermont	80.8	24.3	8.9	86.4	29.4	11.1	91.0	33.1	13.3
Virginia	75.2	24.5	9.1	81.5	29.5	11.8	86.6	34.0	14.1
Washington	83.8	22.9	7.0	87.1	27.7	9.3	89.7	31.0	11.1
West Virginia	66.0	12.3	4.8	75.2	14.8	5.9	82.8	17.3	6.7
Wisconsin	79.6	17.7	5.6	85.1	22.4	7.2	89.8	25.7	8.4
Wyoming	83.0	18.8	5.7	87.9	21.9	7.0	91.8	23.8	7.9

Source: U.S. Census Bureau, 1990 Census of Population, CPH-L-96; 2000 Census of Population, P37, "Sex by Educational Attainment for the Population 25 Years and Over"; 2009 American Community Survey, R1501, "Percent of Persons 25 Years and Over Who Have Completed High School (Includes Equivalency)," R1502, "Percent of Persons 25 Years and Over Who Have Completed a Bachelor's Degree," and R1503, "Percent of Persons 25 Years and Over Who Have Completed an Advanced Degree;" <http://factfinder.census.gov/>, accessed February 2011.

Source: Mullin and O'Brien (2012)

6.3 Method

Choosing the Method

- The first matter is to determine the choice of visualization method
- Not necessarily committing just yet to a specific chart or graph type, though we might have some in mind
- Rather, this is about the general family or collection of chart types as defined by their primary storytelling method
- For example, a bar chart serves the function of comparing categories of values
- A line chart, by contrast, enables us to show changes of values over time
- Geo-spatial data can often be best displayed over a map
- Your choice of visualization method will be mostly driven by the your editorial focus and what you have learned about your data

Classifying Methods

- There are a number of ways of classifying the variety of methods for visualizing data, but here is a suggested taxonomy:
  - Comparing categorical values
  - Assessing hierarchies and part-of-a-whole relationships
  - Showing changes over time
  - Mapping geo-spatial data
  - Charting and graphing relationships
- Of course, there are often overlapping functional or storytelling features inherent to the chart types that sit under these method headings

Bertin's Hierarchy: Discriminate

- The highest level of Bertin's interpretive acts concerned whether we are able to visually discriminate between different data marks or data series
- can we actually see and read the data being presented?
- We must make sure that the way we visually distinguish different categorical and quantitative values is legible and is in no way hidden by way of unnecessary clutter, noise, or distraction

Bertin's Hierarchy: Order or Ranking

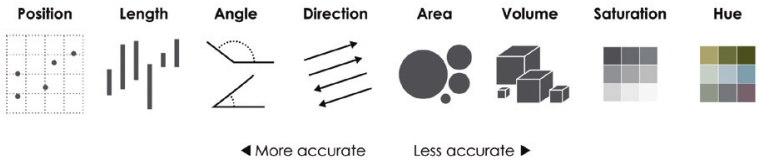
- The second act refers to being able to satisfactorily judge the relative order or ranking of values in terms of their magnitude
- This is basic pattern matching where we seek to determine the general hierarchy of the values being displayed
- where is the most and where is the least, which is the biggest and which is the smallest

Bertin's Hierarchy: Values

- The lowest-level act relates to judging values
- Studies have shown how the effectiveness of different visual variables can be ranked based on which most accurately support comparison and pattern perception

Cleveland and McGill

- Bertin was the first to propose such a hierarchy and his work has been tested, developed, and refined by Cleveland and McGill
- The focus of the study was to determine how accurately people read the visual cues above (excluding shapes), which resulted in a ranked list from most accurate to least accurate

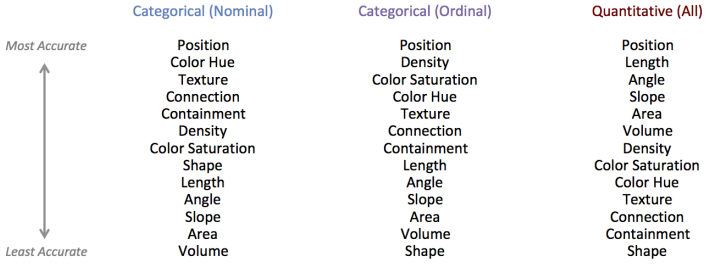


Source: Yau (2013)

Jock MacKinlay

- In the following presentation, we see the most recent version created by MacKinlay
- Each column represents the three main data types
- note that there is no distinction between ratio and interval-scale types of quantitative variables
- Within each column you have an ordering of the most accurate and least accurate visual variables according to their interpretive precision

MacKinlay: Ranking



Source: Kirk (2012)

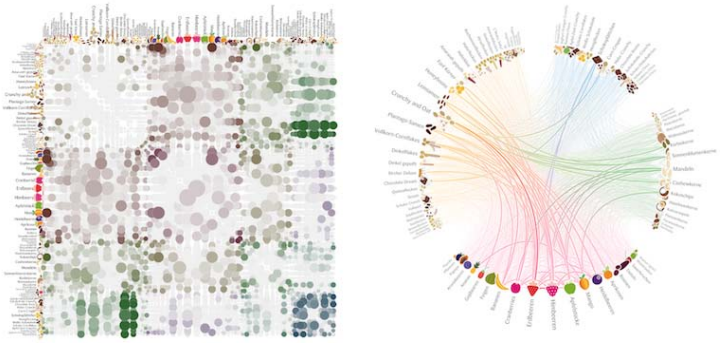
Rankings

- The studies by Bertin, Cleveland and McGill, and then MacKinlay focus on the fact that our visual system isn't capable of absolute measurements
- Therefore, frameworks like this simply propose a guide to understand which variables will be better at delivering relative measurements but with highest accuracy
- In other words, the higher up the column the easier it will be for your reader to accurately interpret values represented by those variables
- Two problems prevent you from always using the "best" choice for every variable
  - Multiple Variables
  - Visual Quality (Art vs. Science)

Metaphor

- Visual metaphors are about integrating a certain visual quality in your work that somehow conveys that extra bit of connection between the data, the design, and the topic
- It goes beyond just the choice of visual variable, though this will have a strong influence
- Deploying the best visual metaphor is something that really requires a strong design instinct and a certain amount of experience.

Example: Muesli



Source: Kirk (2012)

7 Presentation

Presentation

- The presentation of data involves thinking about pretty much every other design feature that might be included in our visualization
- Here, we are determining the following:
  - The use of color
  - The potential of interactive features
  - The explanatory annotation
  - The architecture and arrangement
- The decisions we make about these layers should be focused on delivering extra meaning, intuitiveness, and depth of insight to our readers or users

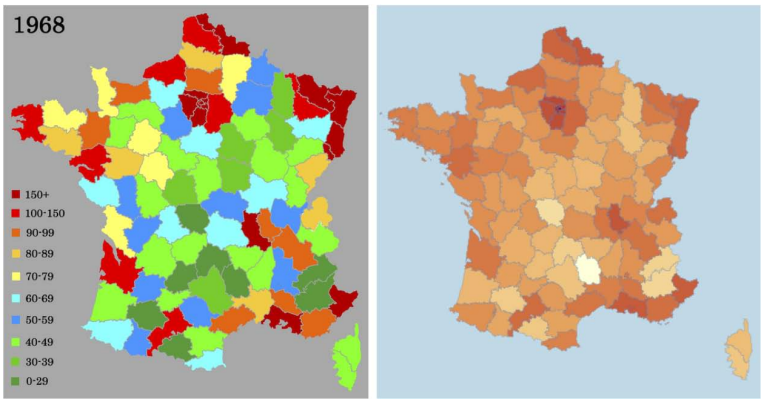
7.1 Color

Use of Color

- Can color be used to represent quantitative data?
- Specifically, when the "hue" property of color is used
- Take a look at this spectrum of colors: if these squares were representing quantitative data, which would be the biggest? How about the smallest?
- Which is bigger, red or blue?

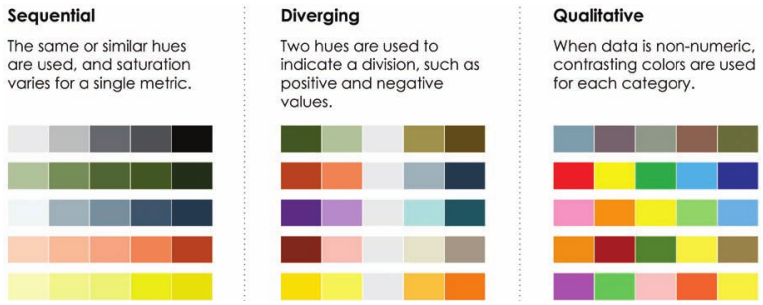
Source: Kirk (2012)

Example: Hue or Saturation



Source: Kirk (2012)

Overview Color Schemes



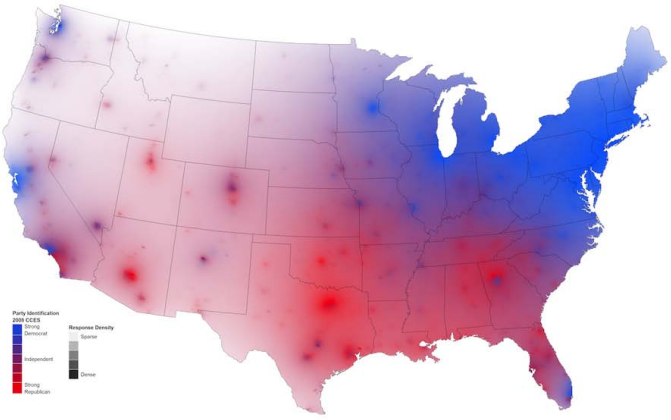
Source: Yau (2013)

Variation

- A narrow color span restricts the amount of difference between shades
- A wider color spanmakes it easier to see differences
- If you don't pay attention to the context of the data, you might show patterns that look obvious but are not significant

Source: Yau (2013)

Utilizing Color Understanding



Source: Kirk (2012)

7.2 Readability

Foregrounding

- To bring the data layer to the fore
- In addition to the representation of data, we also look to employ color to help create visual depth and a sense of hierarchy in our designs
- The clutter that can occur between background presentation and the foreground data representation makes it a real challenge to efficiently establish a sense of visual hierarchy
- The brain and the eyes otherwise have to work especially hard to draw any insight
- What we are trying to establish is a clear sense of the most important signals brought to the foreground and the less important contextual or decorative elements pushed into the background

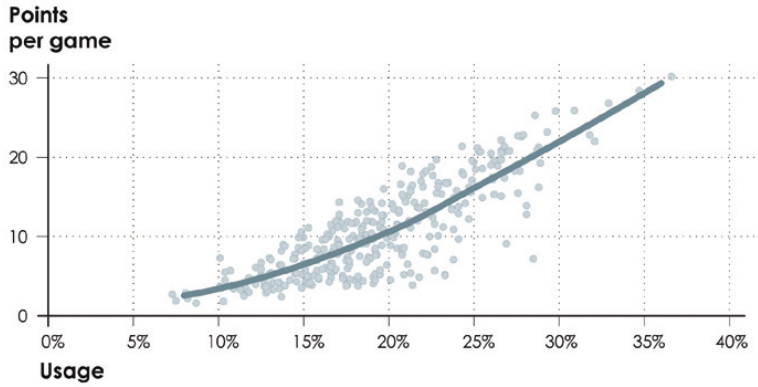
Dampening

- Another important property to take notice of, in the relationship between foreground and background, is the careful deployment of chart apparatus, such as the axes, gridlines, tick marks, borders, titles – any chart property you may use to frame and reference your data
- Don't be afraid to remove or dampen the visible presence of such elements, particularly as the defaults in many tools are set to black
- We are automatically tempted to make things darker, bolder, more prominent, more imprisoned
- Where possible, minimize, dampen, or even remove some of these chart properties because we want to let the data stand out and facilitate our ''seeing'' of its qualities

Visual Hierarchy

- When you look at visualization for the first time, your eyes dart around trying to find a point of interest
- When you look at anything, you tend to spot things that stand out, such as bright colors, shapes that are bigger than the rest, or people who are on the long tail of the height curve
- You can use this to your advantage as you visualize data
- Highlight data with bolder colors than the other visual elements, and lighten or soften other elements so they sit in the background
- Use arrows and lines to direct eyes to the point of interest
- This creates a visual hierarchy

Scatterplot: Foregrounding



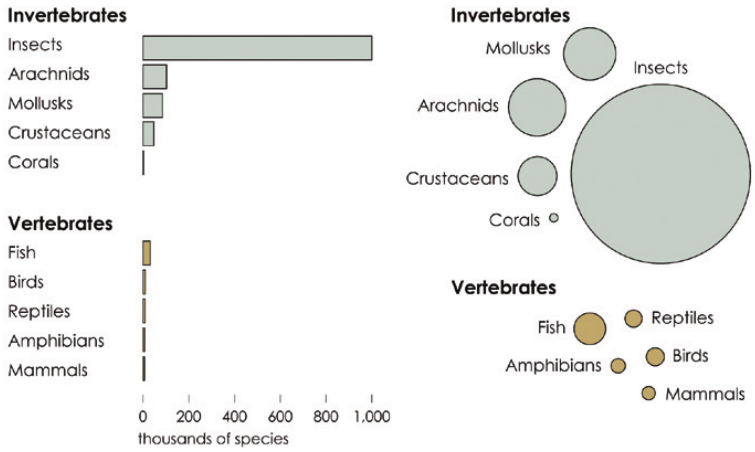
- more descriptive axis labels and less prominent value labels can help

Source: Yau (2013)

Readability: Allow Comparisons

- Allowing comparisons across points is the main purpose of visualizing data
- In table form, you can compare only point by point, so you place data in a visual context to see how big one value is relative to the rest and how all the individual data points relate to each other
- As a way to better understand data, your visualization isn't useful if it doesn't fill this basic requirement
- Even if you just want to show that values are equal across the board, the key is still to allow that comparison and conclusion to be made

Trade-Off: Comparisons & Actual Value



Source: Yau (2013)

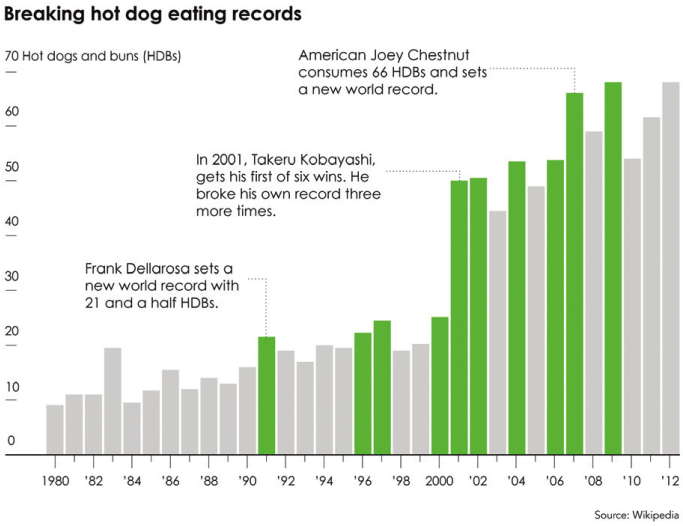
Represent Context

- Context helps readers relate to and understand the data in a visualization better
- It provides a sense of scale and strengthens the connection between abstract geometry and colors to the real world
- You can introduce context through words that surround a chart, such as in a report or story, but you can also incorporate context into the visualizations through your choice of visual cue and design elements

Highlighting

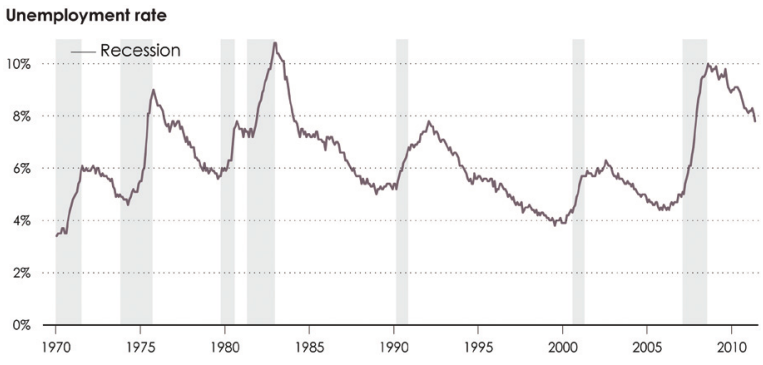
- Highlighting can guide readers through the data and direct eyeballs to the most important parts in a graphic
- It reinforces what people might already see or draw attention to areas or data points that people should see
- To draw visual attention to a data point, you simply do what you would in real life
- You make it stand out. Speak a little louder. Make it a little brighter.
- Edit an area or point in a visualization – while keeping the data, its visual cues, and readability in mind—to differentiate it from the rest
- Use a brighter or bolder color, draw a border, thicken a line, or introduce elements that make the point of interest look different

Example: Color, Labels



Source: Yau (2013)

Example: Background Highlighting



Source: Yau (2013)

7.3 Interactivity

Interactivity: Variables and Parameters

- Manipulating variables and parameters
- The complexity of some data frameworks often means we are trying to find ways of showing many dimensions of stories within a single display or to facilitate different combinations of variables for exploratory visual analysis
- The ability to select, filter, exclude, or modify certain variables is a valuable way of letting the user interact with different slices of the data
- Furthermore, grouping and sorting options are common facilities for extracting new insights

Interactivity: View

- Adjusting the view
- In contrast to manipulating variables, this is more about adjusting the user's lens or window into the subject
- When we have hierarchical or high-resolution data, the ability to perform vertical exploration through the different layers of detail is an important feature
- This can be particularly valuable in map-based visualizations where you may wish to pan around the landscape and zoom through different levels of magnification
- You would see the benefit of this in a project such as the "Wind Map" that we saw earlier, enabling the user to dive into different parts of the country or those areas with strong winds that would be interesting to see in more detail

Interactivity: Annotation

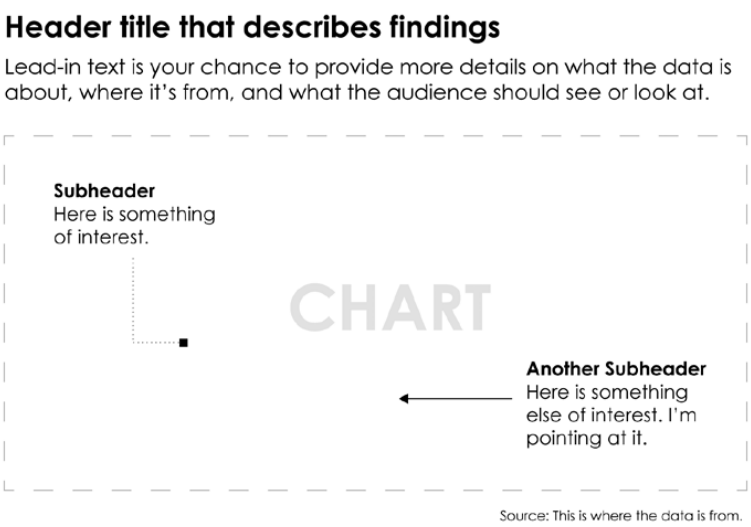
- Annotated details
- in interactive terms, this is about creating extra layers of data detail through interactive events such as hovering or clicking
- This is particularly useful if you want to reveal actual data values or extra detail about a given category or event
- In the earlier section we discussed the degree of accuracy in interpretation and we saw an example of an interactive bubble chart
- As you hovered over the bubbles, you saw a pop-up text display with the raw numbers
- The availability of this type of detail, just a click or hover away from view, might give us greater creative license

Interactivity: Animation

- When we have time-series based data, there is great potential for us to portray our visualization through animation, creating a shifting scene of data as it unravels a compelling data story
- The use of features such as Play, Pause, and Reset can be enhanced by offering manually controllable time sliders (seen in the earlier energy flow example) as well as chapter navigation to skip through key milestones
- The following example below, depicting the expansion of post offices across the U.S. through the years 1700 to 1900 is a perfect demonstration of the potential power of animated data presentation
- While the individual frames are interesting in their own right, the real power of this portrayal comes through the emerging story of the social history of population growth and migration across the country

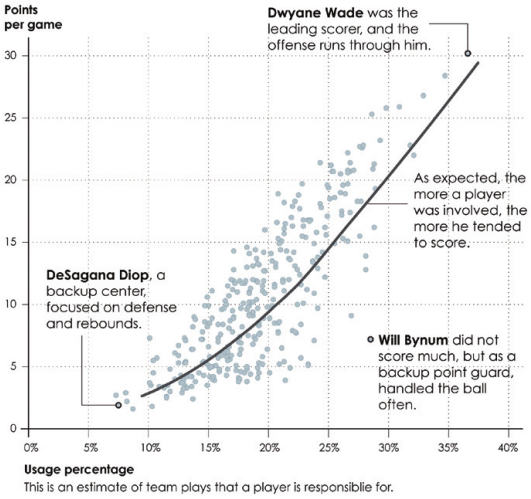
7.4 Annotation

Annotating



Source: Yau (2013)

Annotated Plot

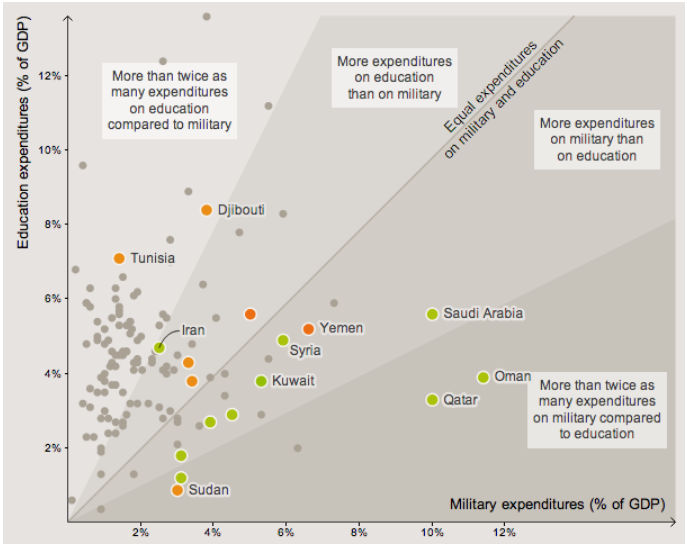


Source: Yau (2013)

Visual Annotation

- Visual annotation: Annotation goes beyond just written explanations and we should consider how to use chart or graphic devices to help draw out important insights visually
- Simple options include features such as gridlines, axes labels, and tick mark
- In an example we have seen earlier, reference lines and background shading is used effectively to help the reader achieve distinction between different tiers of interpretation, as you explore the relationship between what countries spend on education and the military

Example: Visual Annotation



Source: Kirk (2012)

Legends and Keys & Units

- Legends and keys: Always explain the use of color schemes or the varying size of shapes in terms of their categorical or quantitative representation
- Units: You should include details of the units of values being displayed to ensure you don't create ambiguities and potential misinterpretation
- As with many of these annotated features, this is an obvious requirement, something we've had drilled into us since our school days, but you'd be surprised how often they can be left out

Data Sources & Attribution

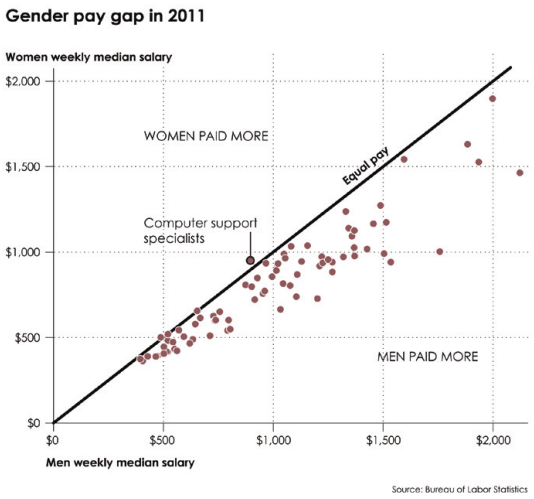
- Data sources: It is vital to include detailed references about from where you have accessed your data or any other sourced element (such as imagery)
- Where you have chance to offer a more detailed narrative, you may wish to explain what treatment you have applied to the data in terms of its quality or analytical transformation
- Attribution: Don't forget to acknowledge those who have either contributed directly, influenced the construction of the design, or those people whose work has acted as a source inspiration

7.5 Math

Explain Statistical Concepts

- If a large proportion of your audience is unfamiliar with statistical concepts, you can annotate to explain or help them relate
- The descriptions in the scatter plot of basketball players are an example
- They don't just point out Dwyane Wade, DeSagana Diop, and Will Bynum
- They also help explain what the corner positions, as well as a partial outlier, on an x-y plot mean so that readers can infer what positions in the middle represent
- The pointer for the trend line is a description of correlation

Example: Gender Pay Gap



Source: Yau (2013)

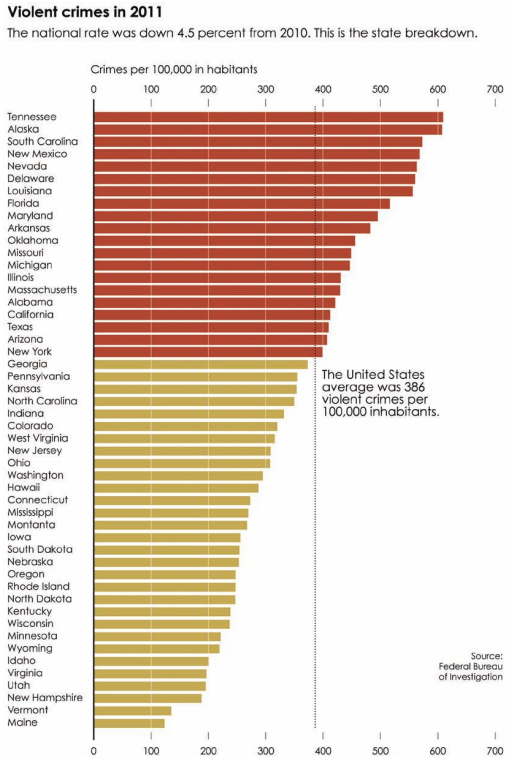
Distributions

- Distributions are another challenging concept
- People have to understand skew, mean, median, and variation, and that observations are aggregated across a continuous value scale when visualized
- For example, it is common for people to interpret the value axis of a histogram as time and the count or density on the vertical axis as a metric of interest
- This leads to confusion, so it is useful to explain the various facets of a distribution

Do the Math

- After you get data, the natural first step is to visualize it directly, but after that, it might be useful to do some math for a different point of view
- This can shift focus toward something more interesting in the data and in some cases, avoid guesswork as readers try to interpret your graphics
- For example, summary statistics, such as mean or median, can serve as a quick point of reference or to provide a sense of scale
- The next figure provides an example

Violent Crimes



Source: Yau (2013)

- Violent crime rates for each state are shown, and bars are colored based on whether they are above the national average
- The distributions of rates isn't especially complex in this example, but it helps you get a sense of where each state lies relative to the national average

More Math

- As an additional step, you can transform the data based on a reference point, rather than just show it in the context of the raw data
- The next Figure shows global gas prices, which you saw previously, relative to average gas price in the United States
- Purple indicates higher gas prices, and green indicates ountries where gas prices were lower
- The two maps show the same data but tell different stories via subtraction and division
- The first map focused on worldwide comparisons, whereas this map provides a simple connection between the data and U.S. readers

Reduce Effort

- The key overall aim is to reduce the amount of work the eye has to undertake to navigate around the design and to decipher the sequence and hierarchy of the display
- For the brain, once again, we're looking to minimize the amount of thinking and ''working out'' that goes on

- We therefore need to carefully consider the choices we make around the size, positioning, grouping, and sorting of all that we show
- As with all visualization design layers, we need to be able to justify the decisions we make about every visible property presented

References

Literatur

Fry, B. (2008). *Visualizing Data – Exploring and Explaining Data with the Processing Environment*. O'Reilly, Sebastopol, CA.

Gerrig, R. J. and Zimbardo, P. G. (2008). *Psychologie, 18. Auflage*. Pearson, München.

Jänicke, H. (2016). Vorlesung visualisierung. online.

Kirk, A. (2012). *Data Visualization – A Successful Design Process*. PACKT Publishing, Birmingham.

Kress, G. and van Leeuwen, T. (2006). *Reading Images – The Grammar of Visual Design, 2nd Edition*. Routledge, London.

Malaka, R., Butz, A., and Hussmann, H. (2009). *Medieninformatik – Eine Einführung*. Pearson Studium, Munich.

Mullin, J. F. and O'Brien, I. R., editors (2012). *Statistical Abstract of the United States: 2012 (131st Edition)*. United States Census Bureau.

Spence, R. (2014). *Information Visualization – An Introduction, 3rd Edition*. Springer, Heidelberg.

Tufte, E. R. (2001). *The Visual Display of Quantitative Information, 2nd Edition*. Graphics Press, Cheshire, Connecticut.

Ware, C. (2004). *Information Visualization – Perception for Design, 2nd Edition*. Morgan Kaufman/Elsevier, San Francisco, USA.

Wegener, R. (2011). *Parameters of context: from theory to model and application*. PhD thesis, Department of Linguistics, Macquarie University.

Wegener, R. (2015). *Continuing Discourse on Language. A functional perspective, Vol. 1*, chapter Studying language in society and society through language: context and multimodal communication. Equinox.

Yau, N. (2013). *Data Points – Visualization that means something*. Wiley.

Zimbardo, P. G., Johnson, R. L., and McCann, V. (2012). *Psychology – Core Concepts, 7th Edition*. Pearson, Boston, USA.