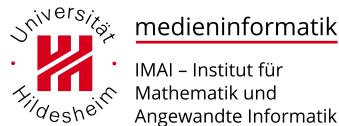


Data

Jörg Cassens

Data and Process Visualization
SoSe 2017



Inhaltsverzeichnis

1	Communicate	1
1.1	Intentions	3
1.2	Process	9
1.3	Representation	11
2	Aspects of Data	21
2.1	Variability	21
2.2	Uncertainty	28
2.3	Context	29
3	Your Data	33
3.1	Data Preparation	34
3.2	Refining Focus	38
3.3	Visual Analysis	40
3.4	Example	42
4	Tutorial	47

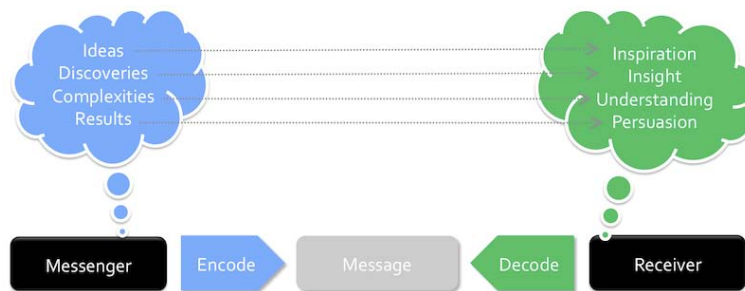
1 Communicate

Data at Heart



ONN: Concentric Circles Emanating From Glowing Red Dot
(2:27)

Communication: Kirk



Source: Kirk (2012)

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

The Reason for Existing I

- Recognizing the trigger behind the project or the origin from where it emerged
- Gives us an idea of the scope and context of what we are about to undertake, how much creative control we might have, whether we've been encouraged to follow a particular creative direction and what ideas have already been formed
- Typically in one of following two ways: you've either been asked to do it or you've decided to do something yourself
- Very different scenarios for working creatively

The Reason for Existing II

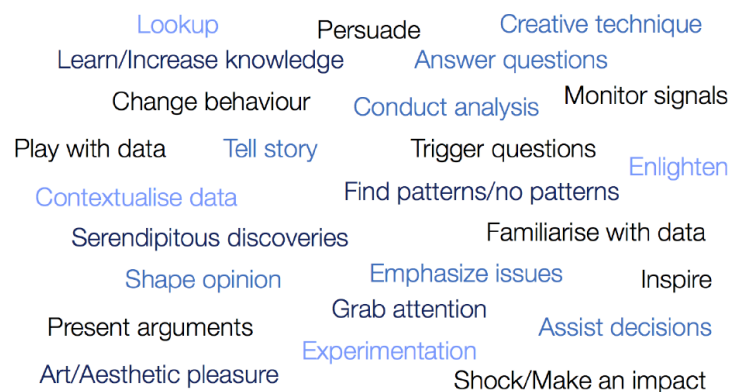
- Asked to do:
 - You will have received or read a brief and possibly had some initial discussions that provided you with an outline of the requirements
 - You might have some instructions and a general idea of what they are seeking
 - From gathering this contextual information, you should have a reasonably clear idea about the background to the project, what you're being asked to do, why you're doing it and who you're doing it for
- Self-initiated:
 - Data sets that you've found about a subject that interests you
 - Test out some theory
 - Completely self-defined, self-determined, and more flexible context than that of a commissioned project

1.1 Intentions

Intended Effect

- Whatever the motivation and background for doing the project, you will inevitably start to form a vision in your mind of what you might be about to create, how it might look, and what it might do
- This is a natural instinct as you embark on a creative process
- This vision might leap into your mind the minute you start to think around the task
- Capture that, take notes (sketches, notes)
- Creating visualizations is a creative process, these early notes can guide our search for an intended effect

Intentions



Source: Kirk (2012)

Dimensions of Intent

- A visualization to assist with the monitoring of signals or facilitating a visual lookup of data will be very different from a design that is intended to grab attention or change behaviour
- Presenting arguments and telling a story is a very different setting to conducting analysis or ‘playing’ with data
- Evidence of different dimensions of intent
- Identifying your intended effect means deciding what you are aiming to achieve and how you are going to achieve it
- At the root of this is an appreciation of your target audience, one of the most important considerations we have to take into account
- During this initial definition and scoping work, it is crucial to profile your intended readers/users

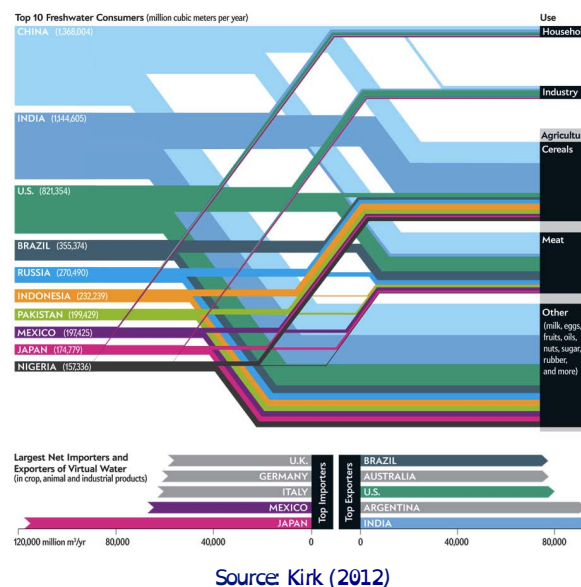
Intent: The Visualization’s Function

- 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

Explain (Example)



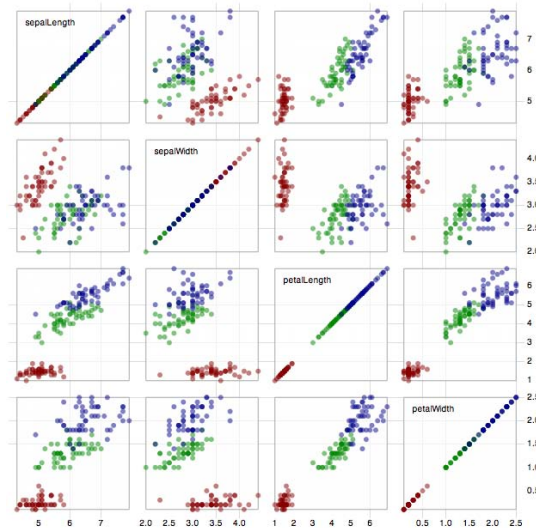
Explore I

- 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

Explore (Example)



Source: Kirk (2012)

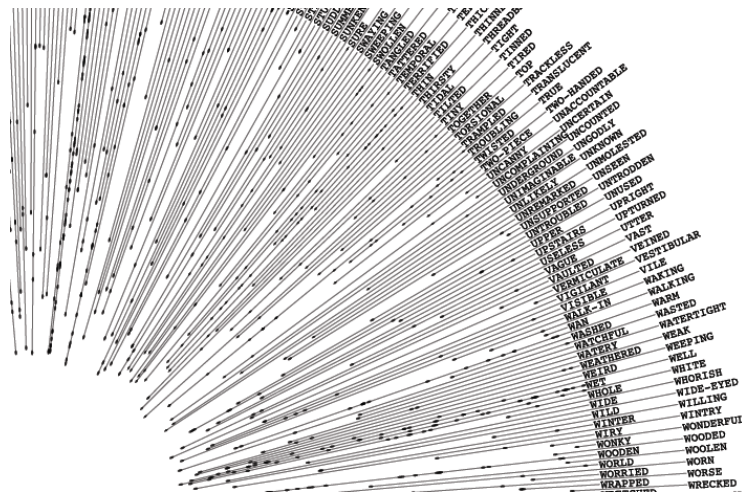
Explore (Interactivity)

- In order to create a truly exploratory experience, interactivity does introduce the potential for so much extra functionality to help immerse the user into a dynamic, problem-solving challenge
- Features such as filtering, sorting, brushing (selecting or isolating certain data values), variable adjustment, and view modification are just some of the important ways you can help a user investigate data
- Also worth highlighting that while explanatory visualization is primarily created for others, exploratory data and the process of visual analysis can be as much for your own discovery purpose as it is for others
- Clearly a particularly relevant function for scientists, for example, to find patterns and unearth key findings in research work before the publication of results

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

Exhibit (Example)



Source: Kirk (2012)

Intent: The Visualization's Tone

- Setting the function is just one part of the “intent” equation
- The clarity of your potential design pathway will be much more apparent as we now consider the second dimension of intent—tone
- Establishing a suitable tone goes beyond function and more towards the style of the design experience
- It concerns the type of stimulus or desired emotional response that you are trying to create
- It is therefore important for you, as the designer, to be able to reason what sort of design will achieve that tone
- Juggling creative and scientific perspectives
- This dynamic poses a significant challenge for any data visualization designer to reason and resolve

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

Leaning

- We need visualizations that look appealing and we need visualizations that perform well
- However, sometimes there has to be mutual recognition that for different scenarios there might be good reason for leaning more towards one direction than the other
 - “We need a chart to help monitor. . .”
 - “We need to present this in a way that persuades people. . .”
- Two situations both aiming to better inform a reader or a user, but the intended effect or outcome from the experience will be different
 - Usability is not User Experience

Learning: Example

- The reaction of a user reading a dashboard full of bar charts and line charts to help monitor monthly performance will be quite analytical and pragmatic in style
 - It is unlikely to involve or stir much emotion
 - The style of the visualization design will be consistent with the intended nature of this particular type of engagement, probably quite sober and with an emphasis on the precision of perception
- Compare that with the intended impact of a presentation that depicts how many lives could be saved if a charity was able to achieve a certain level of fund raising
 - The setting and intent will be more about persuasion making it emotionally charged
 - It will need to attempt to create an experience that is much more personal and more impactful

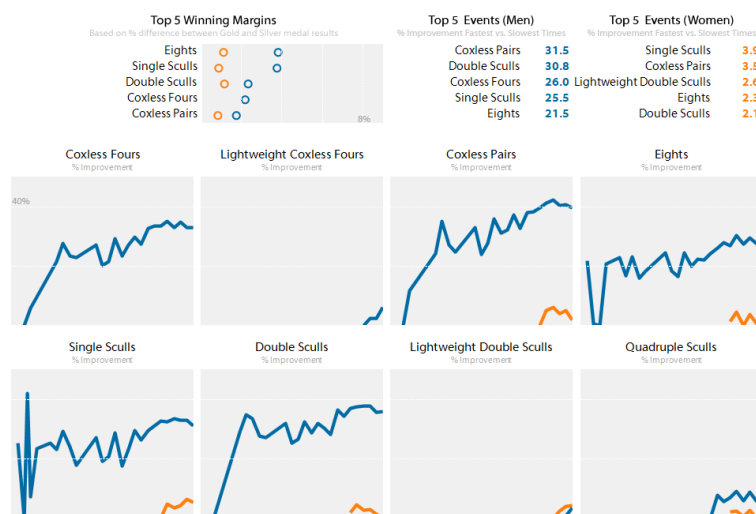
Pragmatic and Analytical

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

Pragmatic and Analytical (Example)



Source: Kirk (2012)

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)).

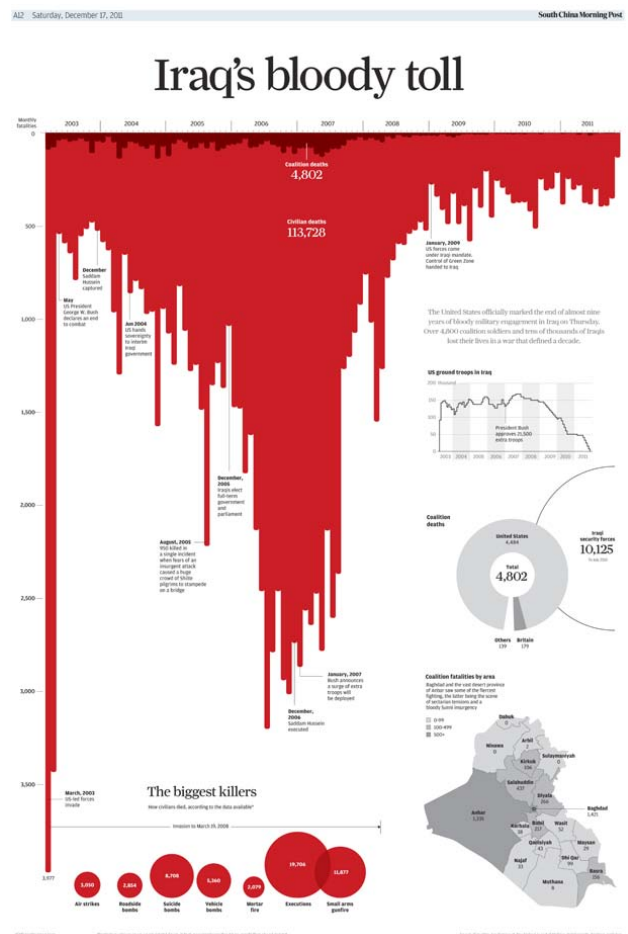
- Abstract visualization, in terms of its tone, is more about creating an aesthetic that portrays a general story or sense of pattern
- You might not be able to pick out every data point or category, but there is enough visual information to give you a feel for the physicality of the data
- This next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

Emotive and abstract (Example)



Source: Kirk (2012)

More Emotion



Source: Kirk (2012)

- For more emotive visualizations, you might be seeking to generate a different type of emotional connection with the design
- Here we see a section taken from a newspaper info graphic that depicted Iraq's blood toll
- Nothing more complex than an upside down bar chart, but the tone is very impactful and metaphorically emphatic

Rehash

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

1.2 Process

Understanding Data

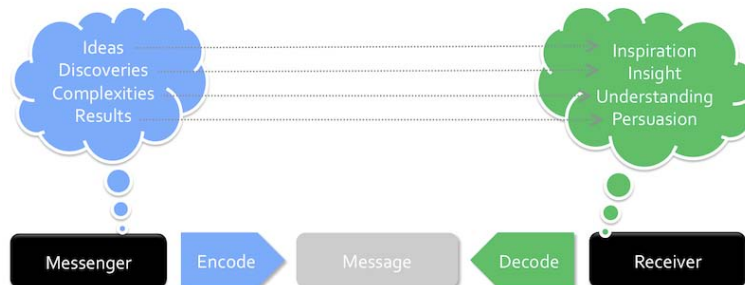
- We looked at *historic examples*, *visual semiotics*, have introduced a *classification framework* for visualizations, and got an understanding of *human visual perception*
- This part of the lecture helps you know your data and what it means to visualize it
- Because of what data represents – people, places, and things – there is always important context attached to the factual numbers
 - Who is the data about?
 - Where is the data from?
 - When was it collected?
- On top of that, most data sets are estimated, so they are *not the absolute truth*; there is uncertainty and variability attached, just like in real life

Rooted in Data

Rooted in Data

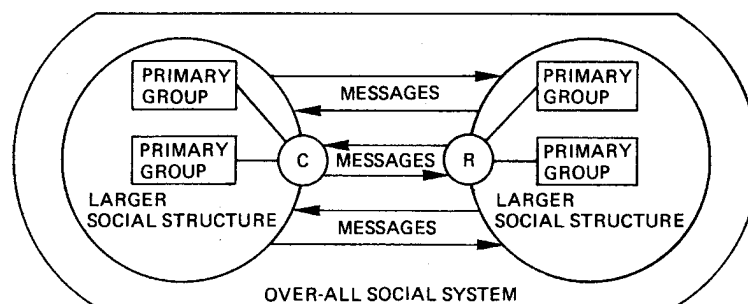
“Visualization is often thought of as an exercise in graphic design or a brute-force computer science problem, but the best work is always rooted in data. To visualize data, you must *understand what it is*, what it *represents in the real world*, and in what *context you should interpret it in*. Data 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)

Communication: Kirk



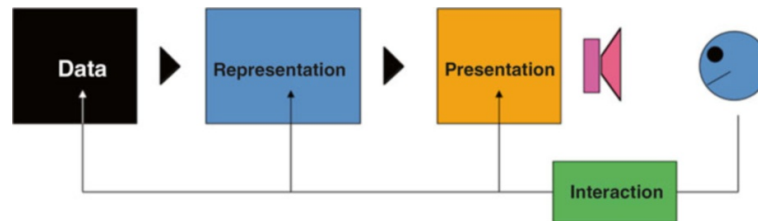
Source: Kirk (2012)

Communication: Riley & Riley



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

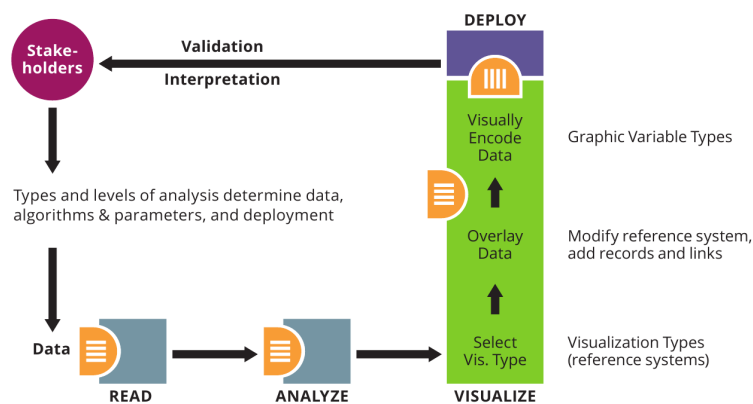
Process: Spence



Source: Spence (2014)

- Work with data
- Representation, i.e. what visualization types to use, interactivity. . .
- Presentation, i.e. human visual system, design. . .

Process: Börner & Polley



Source: Börner and Polley (2014)

Process: Fry

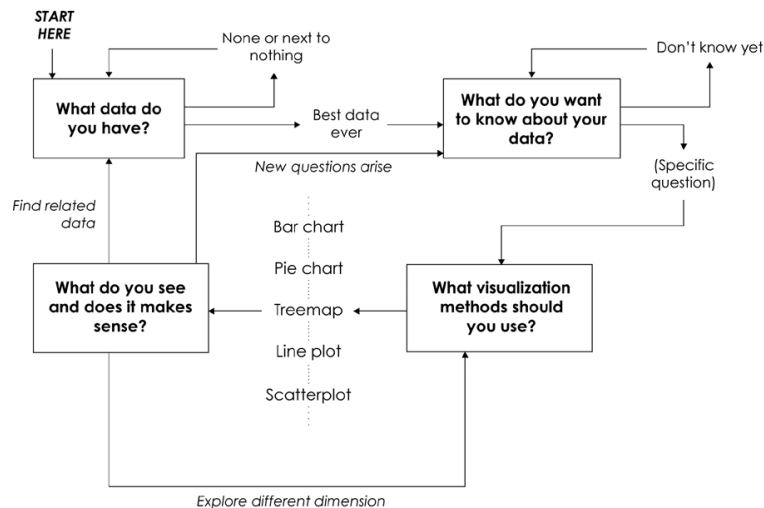
- **Acquire:** Obtain the data, whether from a file on a disk or a source over a network.
- **Parse:** Provide some structure for the data's meaning, and order it into categories.
- **Filter:** Remove all but the data of interest.
- **Mine:** Apply methods from statistics or data mining as a way to discern 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?

Process: Yau



Source: Yau (2013)

Starting Point

- All authors, albeit to a different degree, seem to include working with data early in the process
 - What data do we have (Yau, Fry)
 - What is the context for the data (Yau, Fry, Bömer & Polley)
 - Processing the data (Fry, Bömer & Polley)
- In the following we will take a look at data from a general perspective
 - Data mining, statistics not part of this course
 - We will revisit some aspects when we look at different types of visualizations

Assignment 5.1: Process

- Analyze the different process models
 - Find commonalities and differences
- What role do human aspects play?
- Come up with a unified process model that fits your work flow
- Deliverable:
 - Monday, 26.6., 18:00, learnweb
 - Monday, 26.6., in the course

1.3 Representation

Representation

- Data is more than numbers, and to visualize it, you must know what it represents
- Data represents real life – it's a snapshot of the world in the same way that a photograph captures a moment in time



Source: Yau (2013)

Memories

Context

"If you were to come across this photo, isolated from everything else, and I told you nothing about it, you wouldn't get much out of it. It's just another wedding photo. For me though, it's a happy moment during one of the best days of my life. That's my wife on the left, all dolled up, and me on the right, wearing something other than jeans and a T-shirt for a change. The pastor who is marrying us is my wife's uncle, who added a personal touch to the ceremony, and the guy in the back is a family friend who took it upon himself to record as much as possible, even though we hired a photographer. The flowers and archway came from a local florist about an hour away from the venue, and the wedding took place during early summer in Los Angeles, California." (Yau, 2013)

Data Points

- A lot of information from just one picture, it works the same with data
 - Are pictures not just data as well?
- Single data point can have a *who*, *what*, *when*, *where*, and *why* attached to it, so it becomes meaningful
- Extracting information from a data point isn't as easy as looking at a photo, though
- You can guess what's going on in the photo, but when you make assumptions about data, such as how accurate it is or how it relates to its surroundings, you can end up with a skewed view of what your data actually represents
- You need to look at everything around, find context, and see what your data set looks like as a whole
- When you see the full picture, it's much easier to make better judgments about individual points

Temporal Context

- Imagine I did not tell you more about the wedding photo
- How could you find out more?
- What if you see pictures taken before and after?



Source: Yau (2013)

Context II

- Now you have more than just a moment in time
- You have several moments, and together they represent the part of the wedding when the bride first walked out, the vows, and the tea drinking ceremony with the parents and my grandma, which is customary for Chinese weddings
- Like the first photo, each of these has its own story
- Still though, these are snapshots, and you do not know what happened in between each photo
- For the complete story, you would either need to be there or watch a video
 - Even then, you would still see only the ceremony from a certain number of angles because it is often not feasible to record every single thing

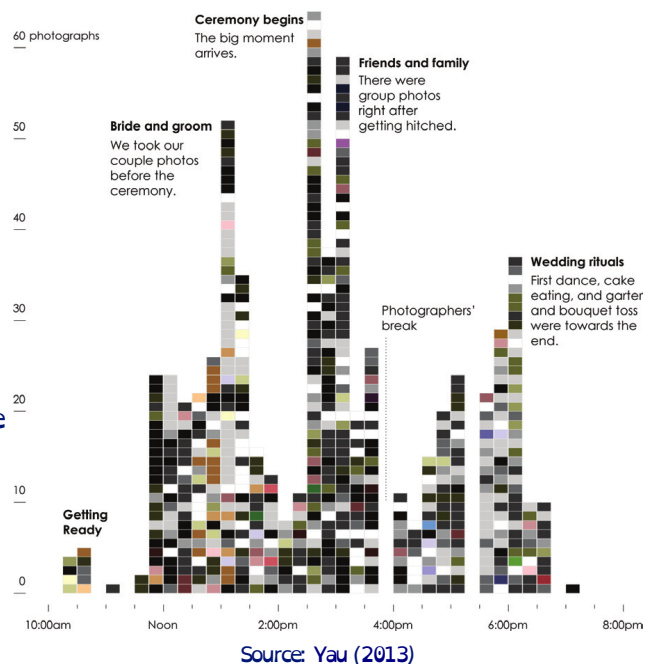
Sampling

- The photos are an abstraction of the real thing
- This is where **sampling** comes in
- It is often not possible to count or record everything because of cost or lack of manpower (or both), so you take bits and pieces, and then you look for patterns and connections to make an educated guess about what your data represents
- The data is a simplification (abstraction) of the real world
- So when you visualize data, you visualize an abstraction of the world, or at least some tiny facet of it
- Visualization is an abstraction of data
- In the end, you end up with an abstraction of an abstraction

Detaching

Wedding colors

Each rectangle represents a photograph during my wedding, and each is filled with the most common color in the picture.



- However, this is not to say that visualization obscures your view
- Visualization can help detach your focus from the individual data points and explore them from a different angle
- This example uses the full wedding data set
- Each rectangle represents a photo from the wedding album; coloured by the most common shade in each photo and organized by time

Time Series

- With a time series layout, you can see the high points of the wedding, when the photographers snapped more shots, and the lulls, when only a few photos were taken
- The peaks in the chart, of course, occur when there is something to take pictures of
- In the grid layout, you might not see this pattern because of the linear presentation
- Everything seems to happen with equal spacing, when actually most pictures were taken during the exciting parts
- You also get a sense of the colours in the wedding at a glance: black for the suits, white for the wedding dress, coral for the flowers and bridesmaids, and green for the trees surrounding the outdoor wedding and reception

Trade-Off

Details vs. Patterns

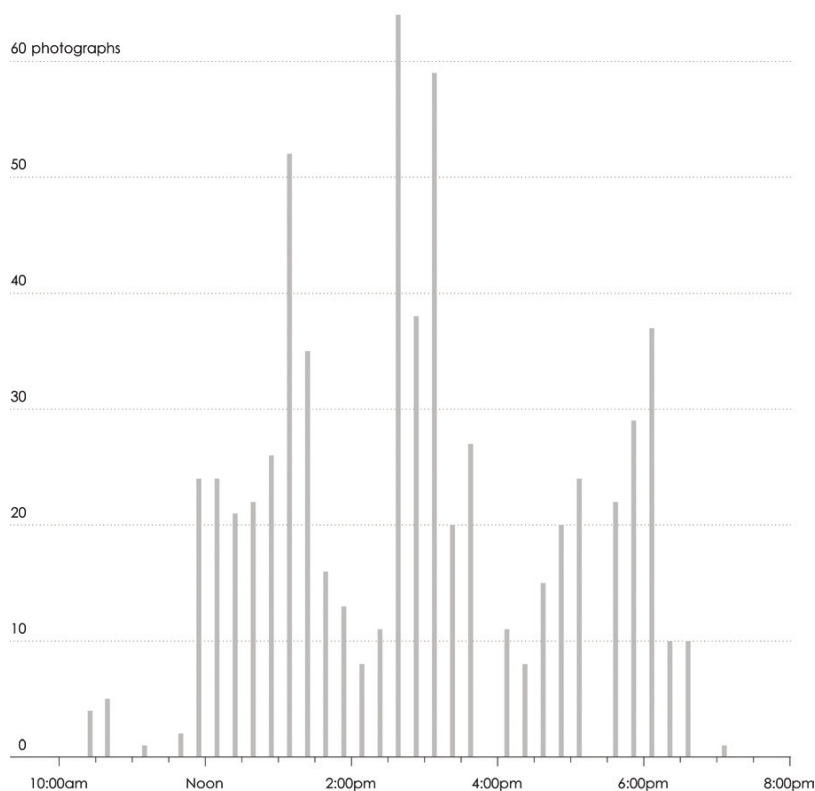
"Do you get the detail that you would from the actual photos? No. But sometimes that level isn't necessary at first. Sometimes you need to see the overall patterns before you zoom in on the details. Sometimes, you don't know that a single data point is worth a look until you see everything else and how it relates to the population." (Yau, 2013)

- Zoom out another level to focus only on the picture-taking volumes, and disregard the colors and individual photos, as shown on the next slide
- It's a bar chart that shows the same highs and lows as the previous figure, but it has a different feel and provides a different message

Abstract Overview

Photographs over time

Our wedding photographers snapped more pictures during the significant events with a peak of 63 during a 15-minute span.



Source: Yau (2013)

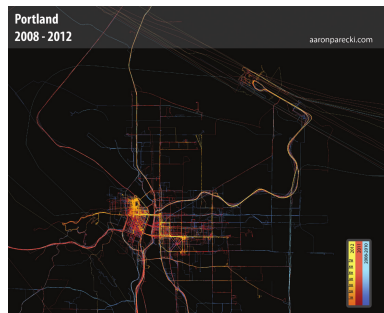
Data and Representation

- The connection between data and what it represents is key to visualization that means something
 - It is key to thoughtful data analysis
 - It is key to a deeper understanding of your data
- Computers do a bulk of the work to turn numbers into shapes and colors, but you must make the connection between data and real life, so that you or the people you make graphics for extract something of value



Connection

- This connection is sometimes hard to see when you look at data on a large scale for thousands of strangers, but it's more obvious when you look at data for an individual
- You can almost relate to that person, even if you have never met him or her
- For example: Portland-based developer Aaron Parecki used his phone to collect 2.5 million GPS points over $3\frac{1}{2}$ years between 2008 and 2012, about one point every 2 to 6 seconds



Source: Yau (2013)

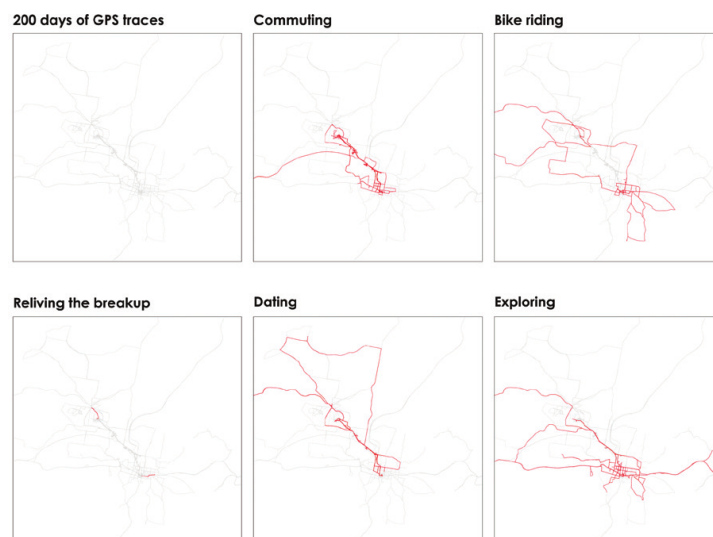
Portland: Meaning

- The map shows a grid of roads and areas where Parecki frequented that are colored more brightly than others
- His housing changed a few times, and you can see his travel patterns change over the years
 - Between 2008 and 2010, shown in blue, travel appears more dispersed
 - By 2012, in yellow, Parecki seems to stay in a couple of tighter pockets
- Without more context it is hard to say anything more because all you see is location, but to Parecki the data is more personal
- It's the footprint of more than 3 years in a city, and because he has access to the raw logs, which have time attached to them, he could also make better decisions based on data, like when he should leave for work.

More Information

- What if there were more information attached to personal time and location data?
- What if along with where you were, you also took notes during or after about what was going on at some given time?
- Artist Tim Clark did this between 2010 and 2011 for his project "Atlas of the Habitual"
- Like Parecki, Clark recorded his location for 200 days with a GPS-enabled device, which spanned approximately 2,000 miles in Bennington, Vermont
- Clark then looked back on his location data and labeled specific trips, people he spent time with, and broke it down by time of year

Atlas of the Habitual

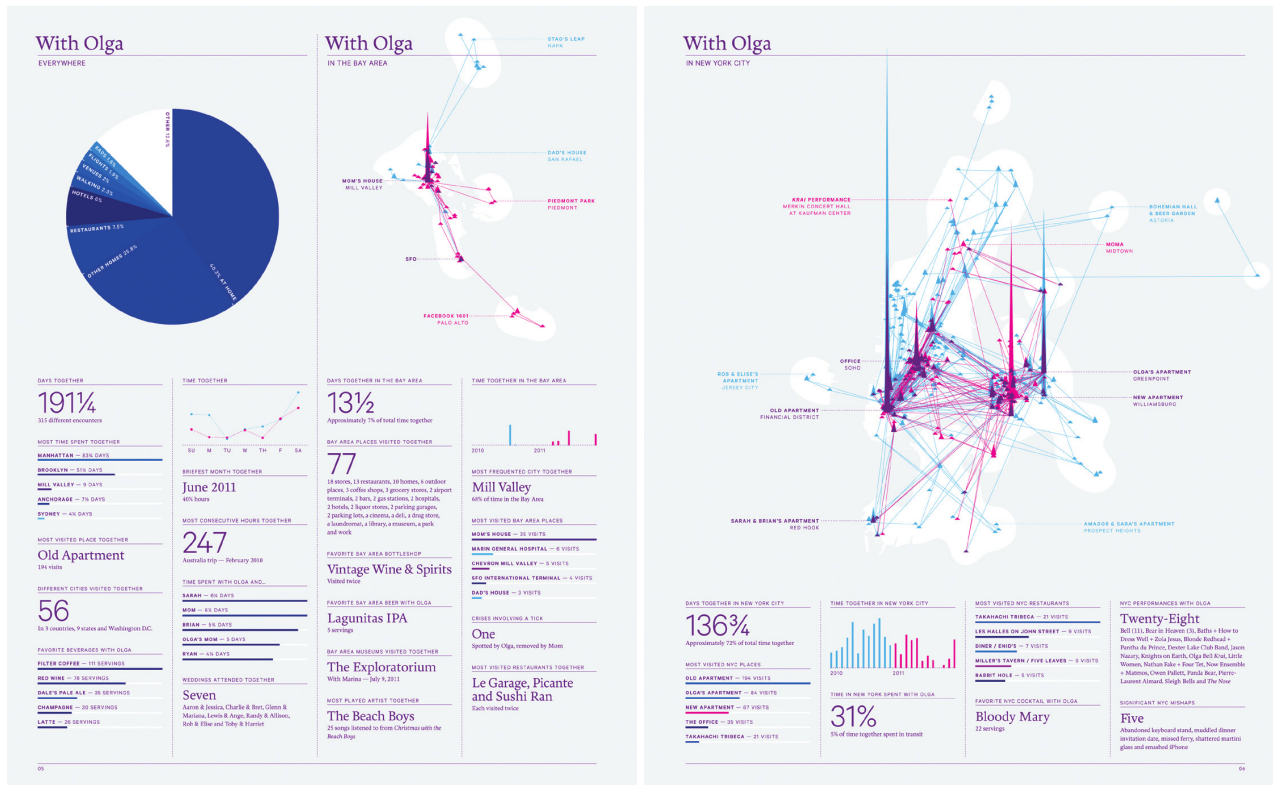


Source: Yau (2013)

Quantified Self

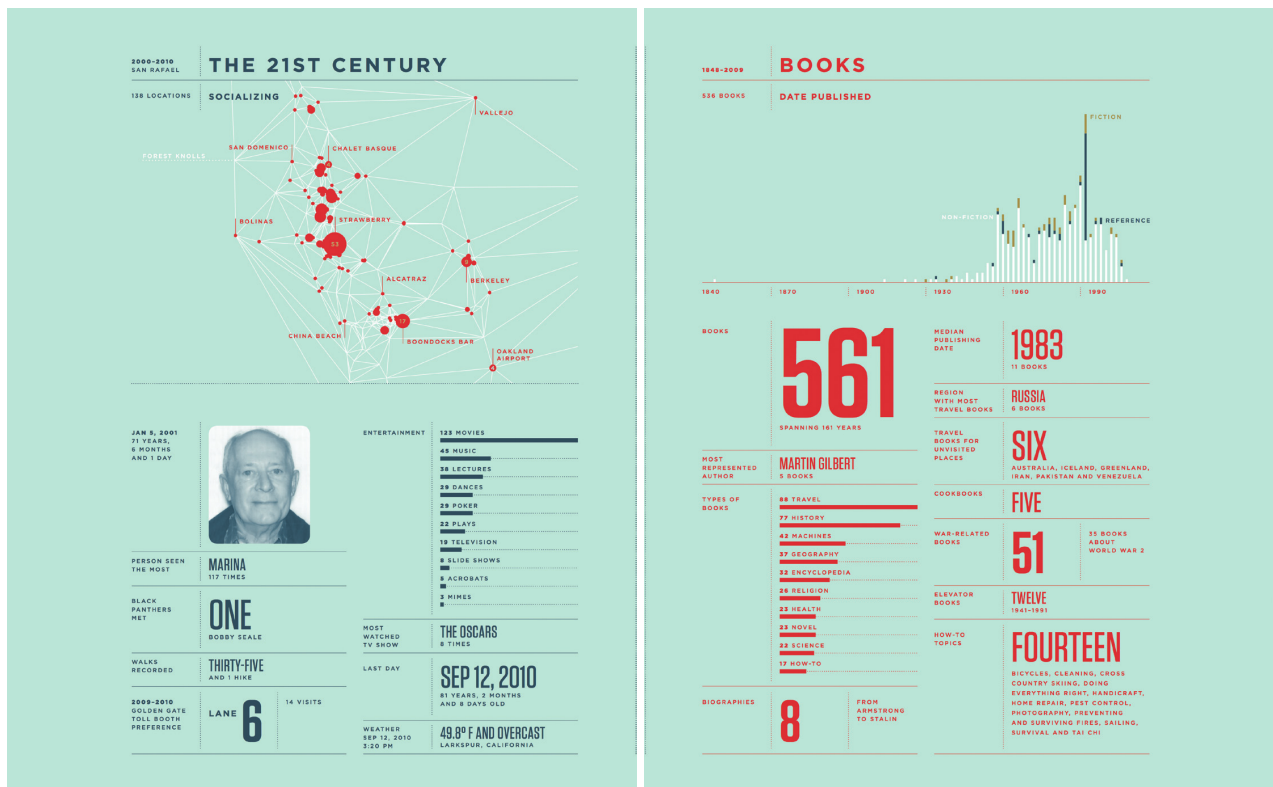
- Nicholas Felton is one of the more well-known people in the “Quantified Self”-area
- Annual reports on himself, which highlight both his design skills and disciplined personal data collection
- He keeps track of not just his location, but also who he spends time with, restaurants he eats at, movies he watches, books he reads, and an array of other things that he reveals each year

Nicholas Felton 2010



Source: Yau (2013)

Nicholas Felton 2010



Source: Yau (2013)

Aggregation

- That data can be valuable to an individual
- What if you look at the data from many individuals?
- The United States Census Bureau collects the official counts of people living in the country every 10 years
- Valuable resource to help officials allocate funds
- And from census to census, the fluctuations in population help you see how people move in the country, changing the neighbourhood composition, how areas grow and shrink
- In short, the data paints a picture of who lives in America
- However, the data, collected and maintained by the government, can show only so much about the individuals, and it is hard to grasp who the people actually are
 - What are their likes and dislikes?
 - What kind of personality do they have?
 - Are there major differences between neighboring cities?

A More Perfect Union

- Media artist Roger Luke DuBois took a different kind of census, via 19 million online dating profiles
- Project “A More Perfect Union”
- You first describe yourself: who you are, where you’re from, and what you’re interested in
- Then, you describe what your ideal mate is like
- DuBois: in the latter, you tell the complete truth, and in the former, you lie
- So when you aggregate people’s online dating profiles, you get some combination of how people see themselves and how they want to be seen
- DuBois categorized online dating profiles by postal code, and then looked for the word that was most characteristic for each area

Personal Sentiment

- Around southern California, where they make the movies, words such as acting, writer, and entertainment appear
- In Washington, DC, words like bureaucrat, partisan appear
- Mostly pertain to professions, but in some areas the words describe personal attributes, favourite things, and major events
- In Louisiana, Cajun and curvy, crawfish, bourbon, and gumbo pop out; but in New Orleans, the most unique word is flood, reflecting Hurricane Katrina in 2005

California



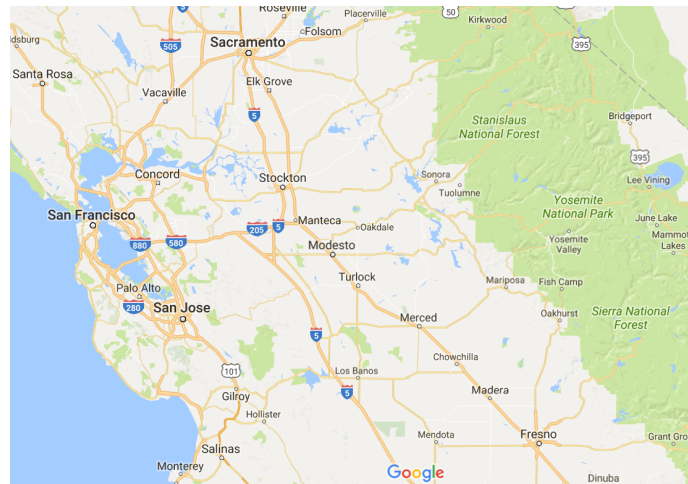
Source: Yau (2013)

California (Detail)

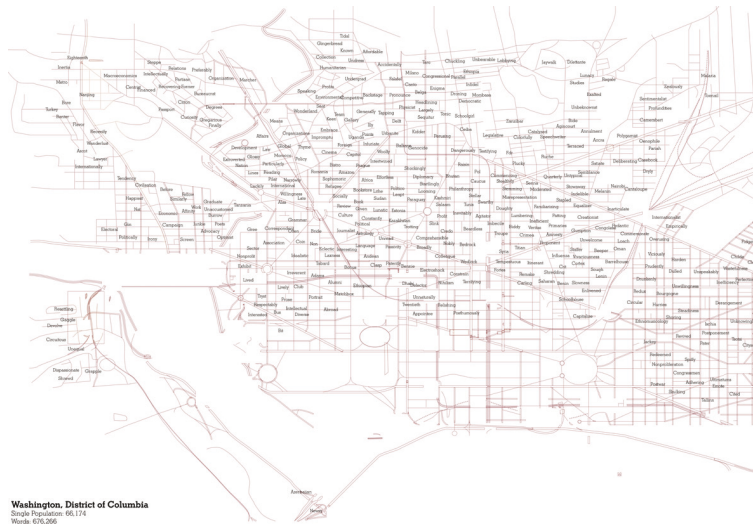


Source: Yau (2013)

California (Detail, Map)



Washington, DC



Louisiana

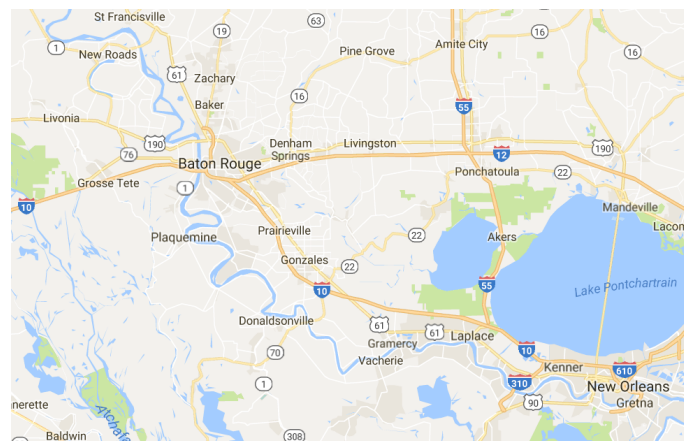


Louisiana (Detail)



Source: Yau (2013)

Louisiana (Detail, Map)



Source: Google Maps

Analysis

- The same sentiment – where data points are recollections and reports are portraits and diaries – is seen in Felton's reports, Clark's atlas, and Parecki's GPS traces
 - Statisticians and developers call this analysis
 - Artists and designers call this storytelling
- For extracting information from data, though – to understand what's in the numbers – analysis and storytelling are one and the same
- Just like what it represents, data can be complex with variability and uncertainty, but consider it all in the right context, and it starts to make sense

2 Aspects of Data

2.1 Variability

Traces

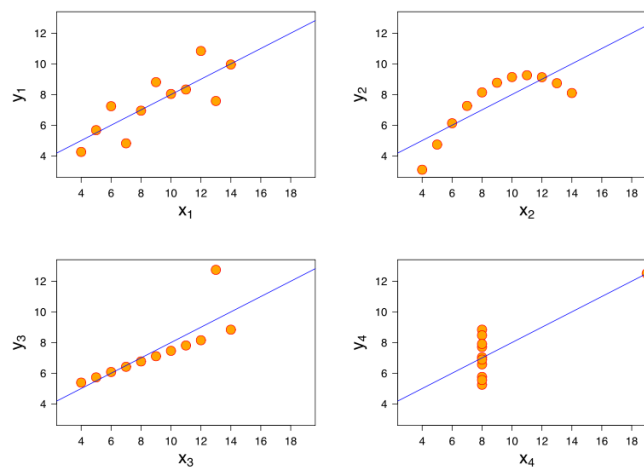


Kristian Cvecek – Tracing fireflies. Source: Yau (2013)

Patterns, Trends and Cycle

- With data, you can find patterns, trends, and cycles, but it's not always (rarely, actually) a smooth path from point A to point B
- Total counts, means, and other aggregate measurements can be interesting, but they're only part of the story, whereas the fluctuations in the data might be the most interesting and important part

Insight?



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

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

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

Data

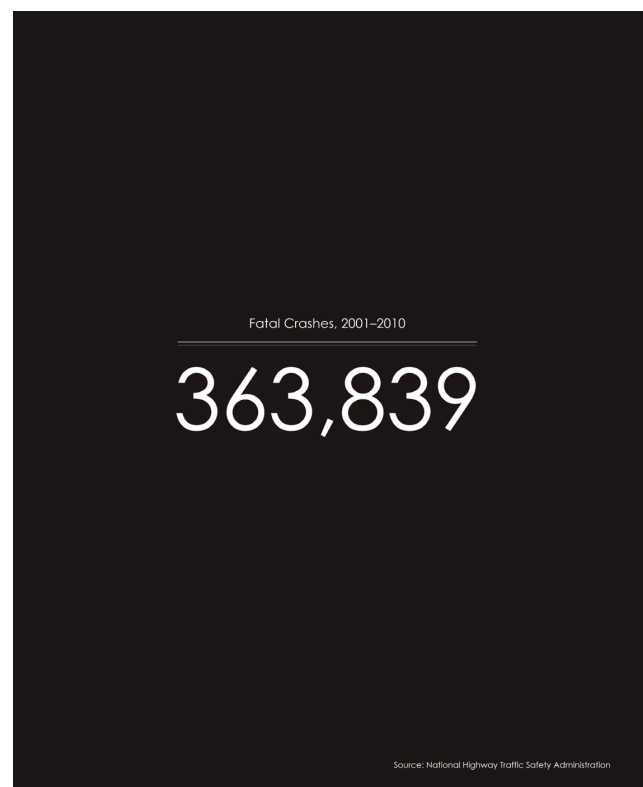
“Anscombe’s quartet” comprises four data sets that have nearly identical simple descriptive statistics, yet appear very different when graphed

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y	4.125	plus/minus 0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively

Sources: Anscombe, “Graphs in Statistical Analysis”,
as reported on [wikipedia](#)

Example: Fatal Crashes

- Between 2001 and 2010, according to the National Highway Traffic Safety Administration, there were 363,839 fatal automobile crashes in the United States
- No doubt this total count, over one-third of a million, carries weight because it represents the lost lives

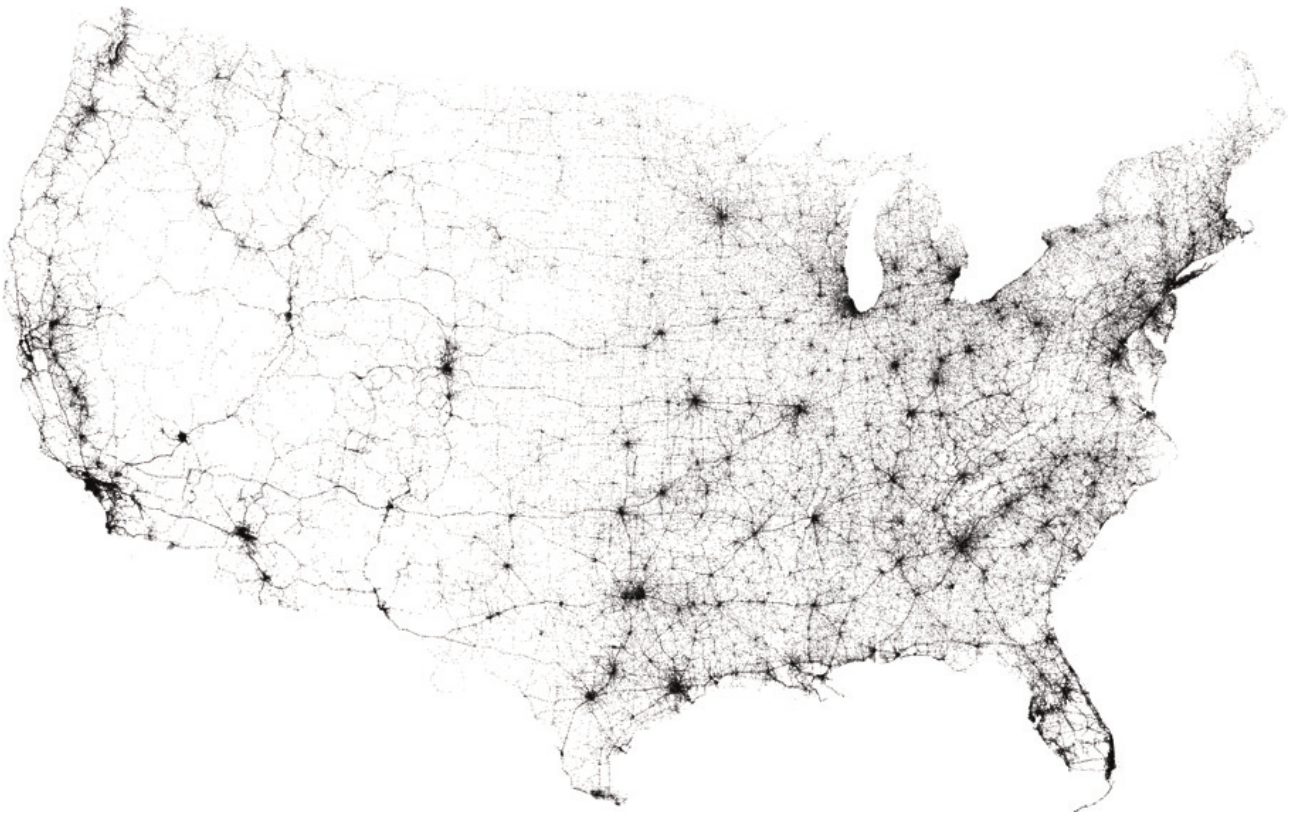


Source: Yau (2013)

Question

- However, is there anything you can learn from the data, other than that you should drive safely?
- The NHTSA provides data down to individual accidents, which includes when and where each occurred
- In the next figure, every fatal crash in the contiguous United States between 2001 and 2010 is mapped
- Each dot represents a crash
- As you might expect, there is a higher concentration of accidents in large cities and major highways; there are fewer accidents where there are fewer people and roads

Location-Mapping

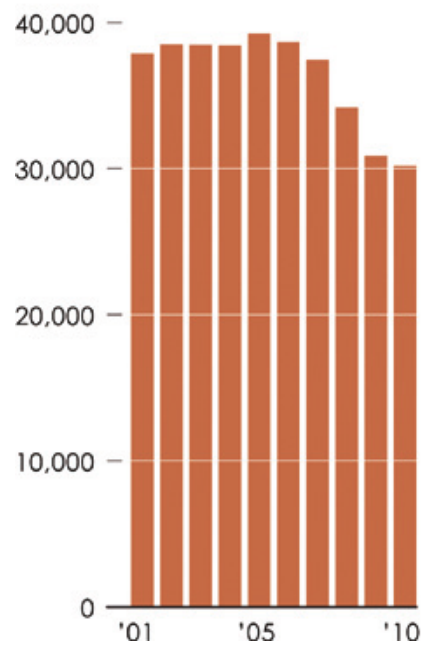


Source: Yau (2013)

Time-Mapping

- 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

Annual fatal crashes



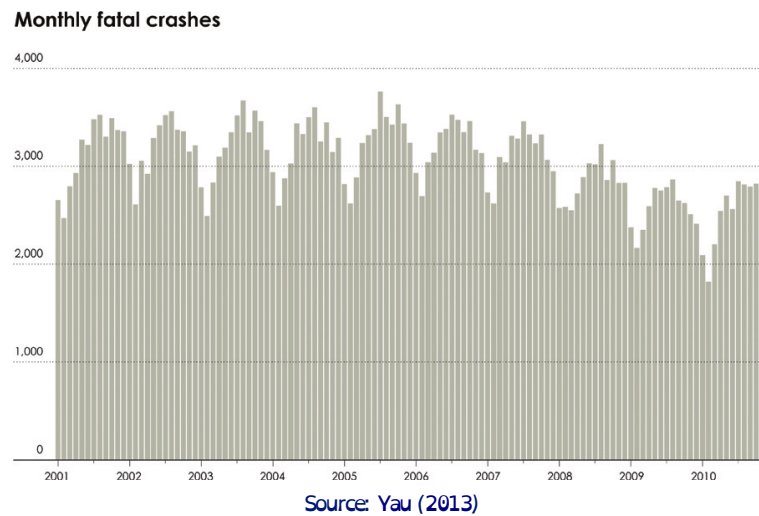
Source: Yau (2013)

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.

Monthly Fatal Crashes



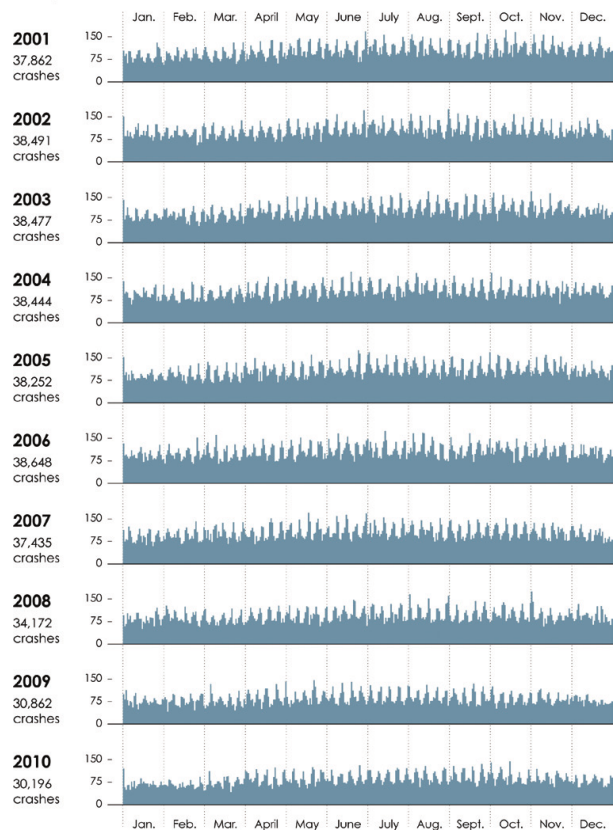
Variability

- However, there's variability when you compare specific months over the years
 - For example, in 2001, the most crashes occurred in August, and there was a small, relative drop the following month
 - The same thing happened in 2002 through 2004
 - However, in 2005 through 2007, July had the most accidents
 - Then it was back to August in 2008 through 2010
 - On the other hand, February, the month with the fewest days had the least accidents every year, with the exception of 2008
- So there are seasonal variations and variation within the seasons
- The question to ask: is this something that should be highlighted?

Daily Crashes

- Go down another level to daily crashes, as shown in the next figure, and you see even higher variability, but it's not all noise
- There still appears to be a pattern of peaks and valleys
- Although it's harder to make out the seasonal patterns, you can see a weekly cycle with more accidents during the weekends than during the middle of the week

Daily fatal crashes



Source: Yau (2013)

Zooming In

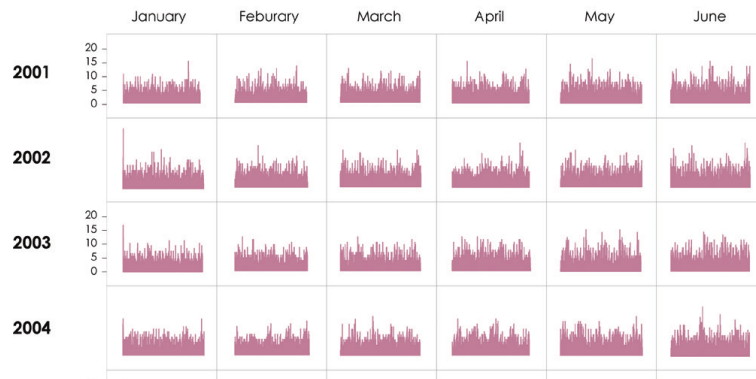
- You can increase granularity to crashes by the hour
- The next figure breaks it down
- Each row represents a year, so each cell in the grid shows an hourly time series for the corresponding month
- With the exception of a new year's spike during the midnight hour, it's hard to make out patterns at 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

Hourly Crashes



Source: Yau (2013)

Hourly Crashes (Detail)



Source: Yau (2013)

Aggregation I

- 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

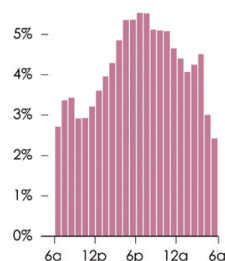
Aggregation II

2001–2010

Fatal crashes by...

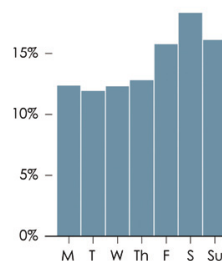
Time of day

Most in the evening and least early morning



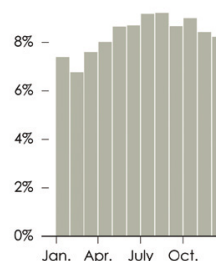
Day of the week

Most on weekends and least middle week



Month

Most in the summer and least during the winter



Source: Yau (2013)

Aggregation III

- There's a small bump in the morning when people commute to work, but most fatal crashes occur in the evening after work
- There are more crashes during the weekend, summed up, it is more obvious
- Finally, you can see the seasonal patterns, but more clearly, with a greater number of accidents during the summer than in the winter

Looking at Data

- The main point is that there's value in looking at the data beyond the mean, median, or total because those measurements tell you only a small part of the story
- A lot of the time, aggregates or values that just tell you where the middle of a distribution is hiding the interesting details that you should actually focus on, for both decision making and storytelling
 - An outlier that stands out from the crowd could be something that you need to fix or pay special attention to

- Maybe the changes over time are a signal that something good (or bad) is happening in your system
- Cycles or regular occurrences could help you prepare for the future
- However, sometimes it isn't helpful to see so much variability; in which case you can dial back the granularity for generalizations and distributions

2.2 Uncertainty

Estimates

- A lot of data is estimates rather than absolute counts
- An analyst considers the evidence (such as a sample), and makes an educated guess about a full population
- That educated guess has uncertainty attached to it
- You do this all the time in your day-to-day – you make a guess based on what you know, read, or what someone told you, and you can say with some (possibly rough) certainty that you're right
- Are you absolutely positive or are you basically clueless? It works the same with data.

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

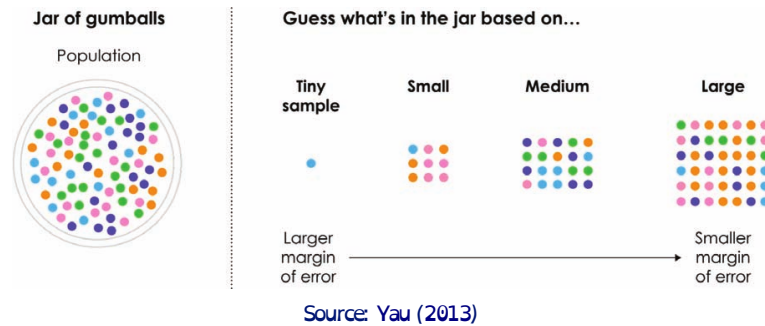
Uncertainty: Census

	Estimate	Margin of error
Total households	114,235,996	± 248,114
Total families	76,254,318	± 230,785
Average family size	3.17	± 0.01
Married-couple family households	56,655,412	± 293,638
Married, 15 and over	50.2%	± 0.2
Divorced, 15 and over	10.5%	± 0.1

- This figure shows estimates about housing
- The margin of error for total households is almost one-quarter of a million

Source: 2010 American Community Survey
Source: Yau (2013)

Sample Size



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

2.3 Context

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

Who Collects

- A quote in a major newspaper carries more weight than one from a celebrity gossip site that has a reputation for stretching the truth
- Similarly, data from a reputable source typically implies better accuracy than a random online poll

Example 2. Gallup, which has measured public opinion since the 1930s, is more reliable than someone experimenting with a small, one-off Twitter sample late at night during a short period of time.

- Whereas the former works to create samples representative of a region, there are unknowns with the latter

Who About

- In addition to who collected the data, who the data is about is also important
- Going back to the gumballs, it's often not financially feasible to collect data about everyone or everything in a population
- Most people don't have time to count and categorize a thousand gumballs, much less a million, so they sample
- The key is to sample evenly across the population so that it is representative of the whole
- Did the data collectors do that?

How

- People often skip methodology because it tends to be complex and for a technical audience, but it's worth getting to know the gist of how the data of interest was collected
- If you're the one who collected the data, then you're good to go, but when you grab a dataset online, provided by someone you've never met, how will you know if it's any good?
- Do you trust it right away, or do you investigate?
- You don't have to know the exact statistical model behind every dataset, but look out for small samples, high margins of error, and unfit assumptions about the subjects, such as indices or rankings that incorporate spotty or unrelated information

What

- Ultimately, you want to know what your data is about, but before you can do that, you should know what surrounds the numbers
- Talk to subject experts, read papers, and study accompanying documentation
- Statistics courses typically teach analysis methods, such as hypothesis testing, regression, and modeling, in a vacuum, because the goal is to learn math and concepts
- For real-world data, information gathering is important
- You shift from
 - “What is in the numbers?” to
 - “What does the data represent in the world; does it make sense; and how does this relate to other data?”
- A major mistake is to treat every dataset the same and use the same canned methods and tools – Don't do that

When


- Most data is linked to time in some way in that it might be a time series, or it's a snapshot from a specific period
- In both cases, you have to know when the data was collected
- An estimate made decades ago does not equate to one in the present
- This seems obvious, but it's a common mistake to take old data and pass it off as new because it's what's available
- Things change, people change, and places change, and so naturally, data changes.

Where

- Things can change across cities, states, and countries just as they do over time
- For example, it's best to avoid global generalizations when the data comes from only a few countries
- The same logic applies to digital locations
- Data from websites, such as Twitter or Facebook, encapsulates the behavior of its users and doesn't necessarily translate to the physical world
- Although the gap between digital and physical continues to shrink, the space between is still evident
- For example, an animated map that represented the "history of the world" based on geotagged Wikipedia, showed popping dots for each entry, in a geographic space

History of the World



 Vimeo: A History of the World in 100 Seconds
(1:40)

Biased History

- The result is impressive, and there is a correlation to the real-life timeline for sure, but it's clear that because Wikipedia content is more prominent in English-speaking countries the map shows more in those areas than anywhere else

Why

- Finally, you must know the reason data was collected, mostly as a sanity check for bias
- Sometimes data is collected, or even fabricated, to serve an agenda, and you should be wary of these cases
- Government and elections might be the first thing that come to mind, but so-called information graphics around the web, filled with keywords and published by sites trying to game Google, have grown up to be a common culprit
- Learn all you can about your data before anything else, and your analysis and visualization will be better for it
- You can then pass what you know on to readers

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

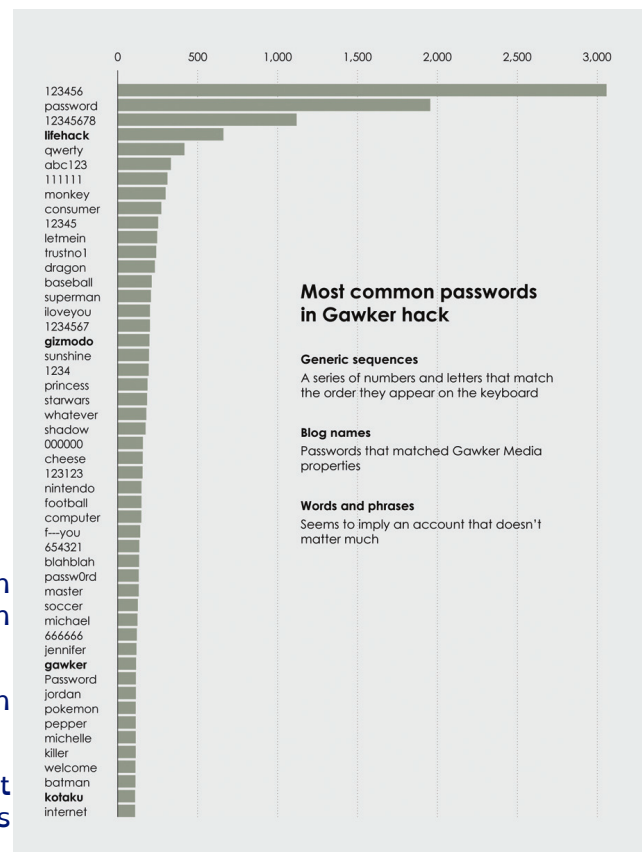
- The passwords were encrypted, but the attackers cracked about 188,000 of them, which exposed more than 91,000 unique passwords
- What would you do with that kind of data?



Gawker.com Gizmodo.com
Lifehacker.com hacked, 1.5 million
usernames/emails/passwords taken

5 minutes ago via web ☆ Favorite 12 Retweet 4 Reply

Gawker Crack



Source: Yau (2013)

- You could highlight usernames with common (poor) passwords, or even create an application that guessed passwords, given a username
- Or you might highlight just the common passwords, as shown here
- This offers some insight into the data without making it too easy to log in with someone else's account

Leaks

- Whether you should use data is not always clear-cut
- For example, on October 22, 2010, Wikileaks, an online organization that releases private documents and media from anonymous sources, released 391,832 United States Army field reports, now known as the Iraq War Logs
 - The reports recorded 66,081 civilian deaths out of 109,000 recorded deaths, between 2004 and 2009
 - The leak exposed incidents of abuse and erroneous reporting, such as civilian deaths classified as "enemy killed in action"
 - On the other hand, it can seem unjustified to publish findings about classified data obtained through less than savory means

Assignment 5.2: Ethical Question

- Should information be free?
- Should everything be free for visualization?
- Do you draw lines?

- Examples:
 - Snowden-Leaks
 - Gawker passwords
 - Iraq-Leaks
- Prepare a short statement about your position for presentation in the course

3 Your 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

Editorial Focus

- Weigh-up the potential appetite of the intended audience – what is it we think they will want to know or will find interesting – and the opportunities that exist within the data – what (data) stories can you find and might portray
- Determining what an audience needs is not always straightforward, particularly when you might have a broad range of different types and background of readers
- Nevertheless, you should still have a sufficiently sympathetic view of how your target demographic will most positively and constructively relate to different slices of analysis of your subject matter
- The design is hugely significant to the success of a project, but without the foundation clarity and justification for the message you are trying to communicate, your resulting visualization will fundamentally lack focus

Selection

- Rather than just throwing everything available at a reader, good visualization involves showing a degree of editorial care
 - You do not have to use all data just because you have it
- This attitude necessary for all types of visualization projects
 - Not only for explanatory pieces
 - Explanatory pieces still have to make data accessible and discoverable

Raw Material

- Data is our raw material, the principle ingredient in our creative recipe
- Irrespective of what we intend or hope to show through our visualization design, the data will ultimately do the talking
- If we don't have the data we want, or the data we do have doesn't tell us what we hoped it would, or the findings we unearth aren't as interesting as we wish them to be there is nothing we can (legitimately) do about it
- An incomplete, error strewn or just plain dull data set will simply contaminate your visualization
- Primary duty is to get on with the task of acquiring our data and immerse ourselves into it to learn about its condition, its characteristics, and the potential stories it contains

3.1 Data Preparation

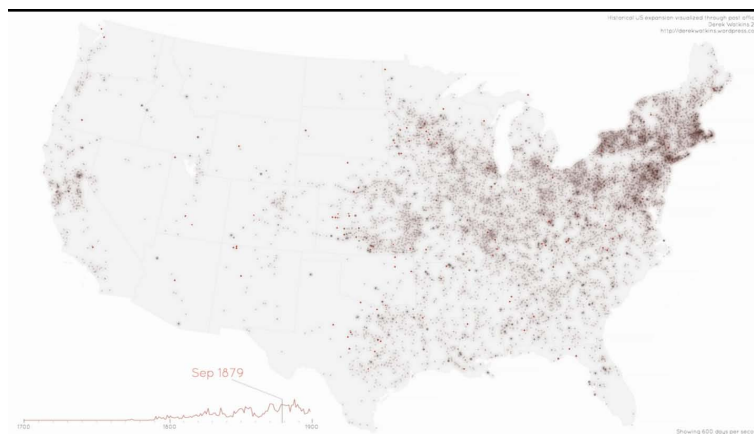
Steps

- 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)

Example: US Postal Service



Source: Kirk (2012)

Example: US Postal Service

- The entire data for this project was scraped from the US Postal Service website
- After cross-referencing the data set with a geographical database to establish accurate geo-locations, ca. 1,500 records (12%) had to be discarded, as not readily “mappable”
- This just shows the great amount of effort and pain that often goes in to sourcing and preparing your data
- No matter from where you are accessing your data, you will often have to work hard to get it into the shape and form that you need it
- Therefore, you need to ensure you have factored in as much time as possible for this vital stage of the process

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: $\neq < > \leq \geq -$ <i>Like a geometric point. Cannot compare directly. Only differences (i.e. intervals) may be compared.</i>
Q – Ratio (location of 0 fixed)	Measurements: Length, Temp, . . . Counts and amounts Operations: $\neq < > \leq \geq - \div$ <i>Like a geometric vector, origin is meaningful.</i>

Data types: Semantics

- Besides their mathematical properties, different data types can carry different inherent semantics
- For example, both Countries and Fruits are of Nominal type
- Only one of them is suitable for a representation on a map of the world
- For this reason, data such as location or time deserve special attention when visualizing them
- Do also make a note of the range of values or at least a sample of the data held against each field

Data types: Examples

Data	Types	Range
Event	Quantitative (interval)	27 different years (1896–2012)
Medal	Categorical ordinal	Gold, silver, bronze
Athlete	Categorical nominal	1500+ different athlete names
Result	Quantitative (ratio)	Race results (9.59s > 4:02:59)
Country	Categorical nominal	96 different country names

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 looking 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

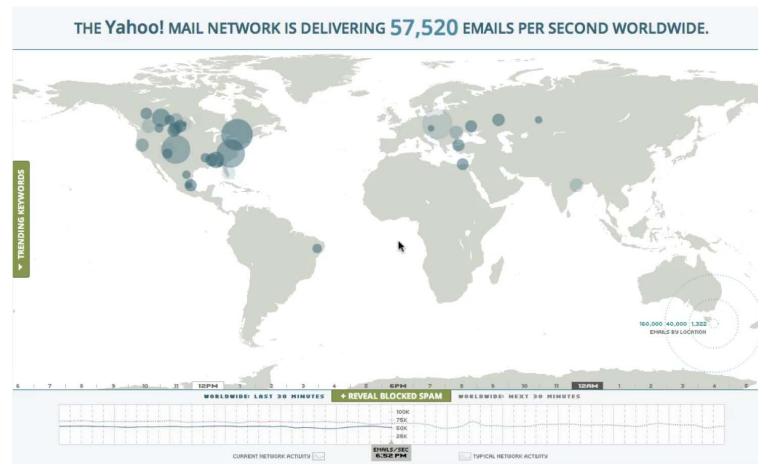
- Here, we focus on preparing and refining it in anticipation of its intended use for analysis and presentation
 - Parsing (split up) any variables, such as extracting year from a date value
 - Merging variables to form new ones, such as creating a whole name out of title, forename, and surname
 - Converting 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)

Excursion: Transformation for Analysis

- As seen before, it is important determine what level of resolution you might wish to, or indeed need to, present your data.
- The decision you take about this may require you to aggregate or disaggregate your data

Example 3. Design agency Periscope were faced with some intricate resolution decisions in their preparatory work for this near real-time visualization developed about the Yahoo! Mail network. The objective was to show the huge volumes of e-mails being sent and processed around the world at any given point, and the efforts Yahoo! is taking to help reduce and intercept spam e-mails.

Yahoo! Mail



Source: Kirk (2012)

Yahoo! Level of Detail

- With approximately 5.6 billion e-mails (plus 20.5 billion spam) sent every day, the sheer amount of data potentially being fed into this project clearly posed a challenge in terms of what level of detail they could reasonably show
- Not just a matter of how they could handle the velocity and volume of new data on the technical side but also what was the appropriate resolution with which to tell this story
 - Headline statistics shown in the titles and presented across a range of supplementary graphics across the project would be representative of the full data quantities
 - For the geo-spatial view, a representative sample of data was extracted
 - Adequate to capture the nuances of the activity seen with the full data set and avoided the technical impracticalities involved in attempting to show 100 percent of the data
 - Geographical data was clustered to a city or regional aggregate, represented by the circle positions and sizes

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

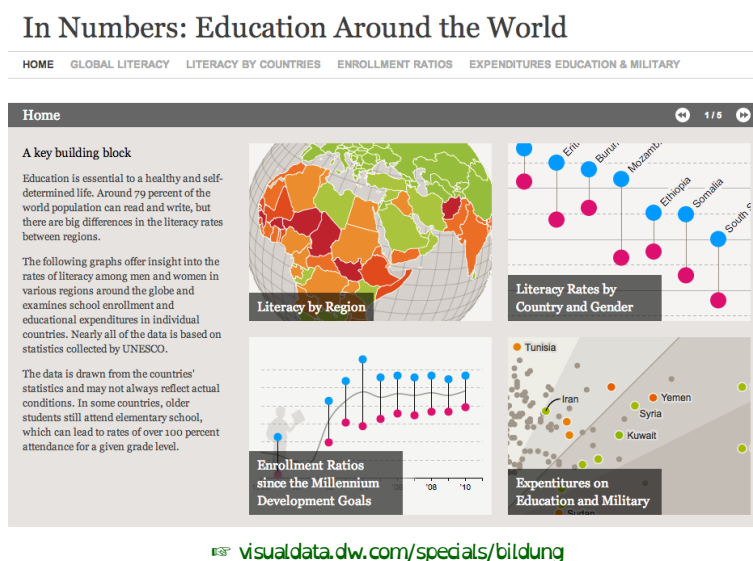
- When you originally access your data, you will likely believe, or hope that you have everything you need
- However, it may be that after the examination and preparation work, you identify certain gaps in your subject matter
- Additional layers of data may be required to be combined (“mashed-up”) with 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

Synthesized Data

- The steps outlined above focus on sampled data
- You might also deal with synthesized data
 - Data from simulations
 - To a degree, data from sensors
- While some of the issues outlined go away with synthetic data, it is helpful to consider the issues as you would with sampled data
- In particular, in synthesized data, you might be able to acquire additional data by re-phrasing the simulation
- Sensed data might be enhanced by additional or better sensors

3.2 Refining Focus

Example: Education Around the World I



Editorial Focus

- Do not be prematurely tempted into diving into the construction of a visualization design, we first need to do more work to fine-tune our analysis of the important messages
- Example: In Numbers: Education Around the World
- A myriad ways of telling data stories about global education matters
- Scoping and definition of the chosen narrative and slices of analysis matter
- Rather than bombarding us with endless pages of facts and figures, or offering seemingly infinite combinations of interactive variable selections, the subject is framed for us around a small number of interesting angles
 - literacy by region
 - literacy rates by country/gender
 - enrollment ratios
 - expenditure on education versus military

Example: Education Around the World II



Story

- Scatter plot of education spend versus military spend for all countries
- Designer takes responsibility for telling the story, providing effective written (labeling and captions), and visual annotation (reference lines and background shading) to help maximize the potential insights
- Inclusion of filtering features to highlight particular countries and regions introduces an exploratory dimension to enable the discovery of further layers of understanding
- A visualization that answers “data questions”: lines of interrogation and the dimensions of interpretation users will likely seek to pursue when reading a visualization design
- Making specific insights accessible – Be able to respond to the most likely and relevant questions a user will raise about the data and the subject matter

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
- The scatter plot we just saw was the perfect method of comparing two quantitative values for all those different countries

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

3.3 Visual Analysis

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

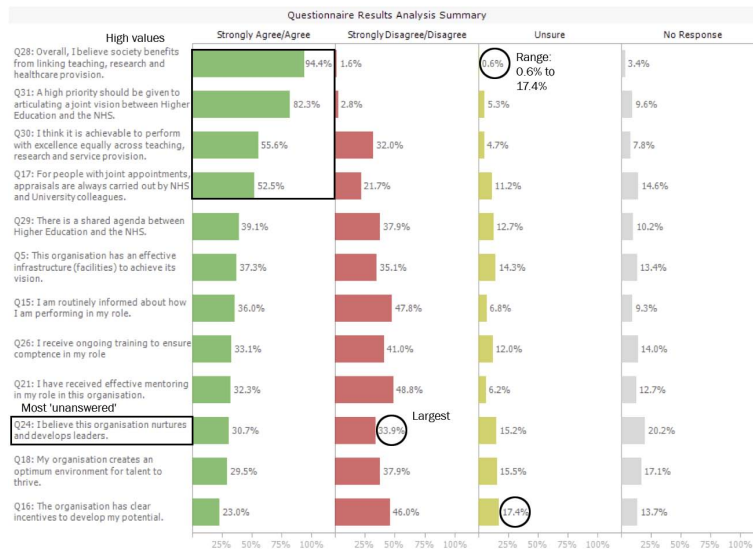
Exploration Dimensions

- 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

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

Example: Comparisons and Proportions

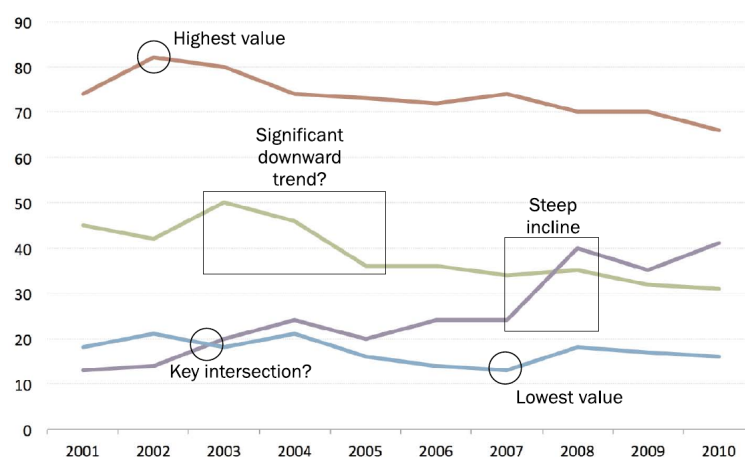


Source: Kirk (2012)

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?

Example: Trends and Patterns



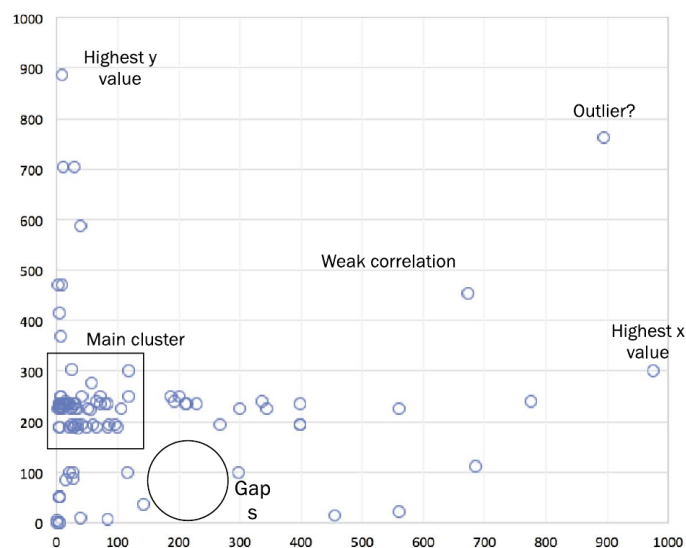
Source: Kirk (2012)

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: Relationships and Connections



Source: Kirk (2012)

3.4 Example

Example

- Take the following sample table of data
- The subject matter is the Olympic games and specifically the total medals won by the top eight participating nations over five recent events
- The selection of the top eight is based on them being the top ranked countries at the Beijing Olympics in 2008
- Suppose you were briefed to unearth some key stories around Olympics medal winning trends in recent years, how would you go about it?

Data

Country	Total medals won in the Summer Olympics				
	2008	2004	2000	1996	1992
United States of America	110	103	92	101	108
People's Republic of China	100	63	59	50	54
Russian Federation	72	92	88	63	112*
Great Britain	47	30	28	15	20
Australia	46	49	58	41	27
Germany	41	49	56	65	82
France	40	33	38	37	29
Republic of Korea	31	30	28	27	29
ALL	951	929	925	842	815

* When part of former Soviet Union. Data from <http://www.databaseolympics.com/index.htm>

Source: Kirk (2012)

Does Anything stand out?

Issues

- *Examination*: Main data issue appears to be that the Russian Federation medals total for 1992 was actually when it was known as the Soviet Union
- Noticeably higher than for all the other Olympic events, due to the contributions of additional member states that then made up the Soviet Union but who are now independent countries
- It will be hard to unpick this value to isolate just those athletes who would now be considered part of the Russian Federation, therefore, it will be sensible to just ignore this value from our analysis
 - Otherwise, it will skew our interpretations

Descriptive

- We can see that the event order goes from left to right in reverse chronological order and the vertical sorting is organized by the most successful nations as at 2008
- In addition to the medal winning totals for the selected countries, we also have the aggregate of all medals across all countries
- We now continue our examination by noting some of the data set's descriptive and statistical properties to develop an increased level of familiarity

Characteristics

- Two variables: Country and event year
- Data Types:
 - Country is a categorical nominal variable with nine values (each country and the aggregate)
 - Event year is a quantitative (interval-scale) variable with five values
 - The maximum country medal count value is 110 medals, the minimum is 15
 - The maximum aggregate value is 951 and the minimum is 815 (but that includes the Russian Federation contribution)
 - Each event year is spaced 4 years apart
 - The longest country name is People's Republic of China, the shortest is France

Preparation

- Gives us a sense of the physicality of the data and the potential influencing attributes that might shape our visualization architecture
- What other data preparation tasks might we undertake?
 - No real *transformation for quality* to undertake in terms of addressing data quality aside from ignoring the Russian Federation total
 - For *transforming for analysis* we may decide to create some calculations to show the percentage of medals won out of each event total
 - You may also decide to abbreviate some of the county values
 - *Data consolidation*: For the purpose of this demonstration, we are going to stick to our original data set on its own but there could be many different options to enhance and contextualize this subject matter

Enhance & Contextualize

- Details behind medal totals (how many golds, silvers, and bronzes)
- Full data set of medal statistics for all the other countries who have competed, not just the recent top eight
- Full data set of medal statistics for every Olympic games
- The number of competitors who were taking part in the games for each country, in order to understand the percentage of success of each team

- Population figures to contextualize the achievements, maybe even sporting participation figures if they were recorded
- Historical milestones of socio-political and geo-political issues
- National flags' image files or URLs to national Olympic associations

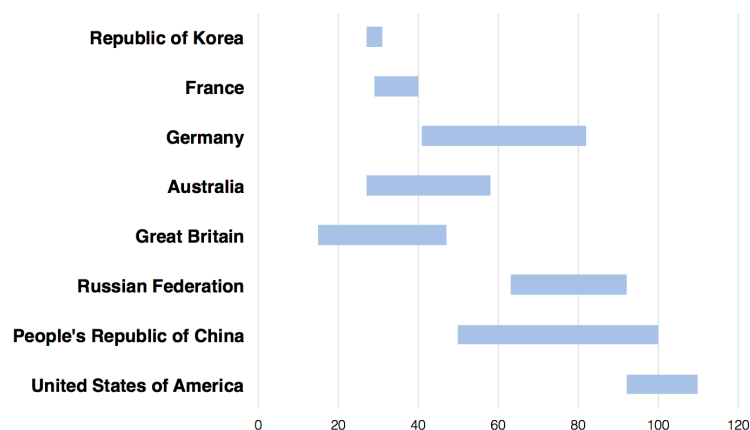
Assignment 5.3: Editorial Focus

- What initial sparks of curiosity crossed our minds when we were given the brief and initially saw the data?
- What dimensions of analysis do we think might be of interest or relevance about this subject matter?
- What data questions will we seek to answer in our visualization design?

Visual Analysis

- To refine our focus we need to commence our visual analysis work to explore our data set and see what comparisons, trends, patterns, and relationships we can identify
- Out of this we will hope to unearth some interesting stories to tell
- Given we have a small data set with only two variables we shouldn't need to embark on too much varied visual analysis
- The first graphic takes a look at the variation of medal winning across the years, showing the range of totals for each country using a floating bar chart

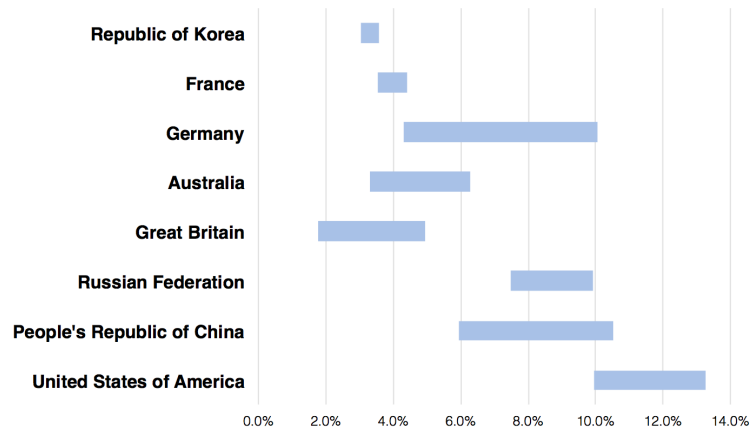
Floating Bar Chart



Source: Kirk (2012)

- Which countries have experienced a significant change in their medal-winning performance levels?
 - We're looking for the widest bars to show the variability, this could be improvement, decline or inconsistency
 - We would identify the spread of Germany and China as being particularly interesting
- Which countries have maintained consistency in their performance levels?
 - Now we're looking for the narrowest bars, the tightest of value ranges
 - This leads to noticing the USA, France, and especially Republic of Korea
- What have been the most interesting country stories in terms of the transition of their performance and rankings?
 - Possibly too hard to see with this chart, but there is potentially something going on with the bars that intersect and exceed the lengths of others
 - At this stage, the story of China seems to stand out as being something to look out for

Normalized Floating Bar Chart



Source: Kirk (2012)

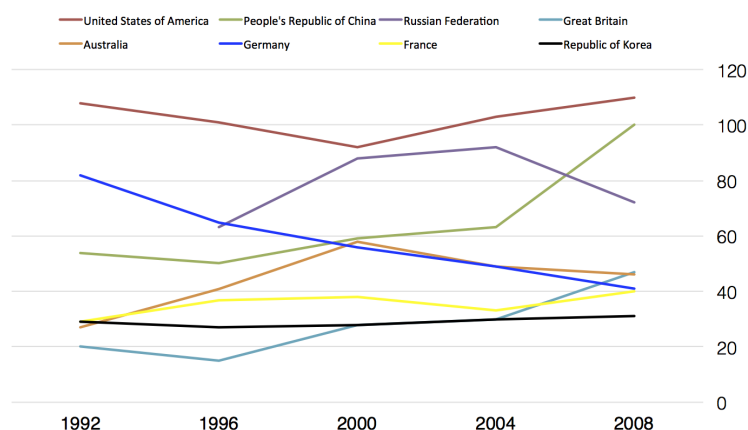
Effects of Normalization

- Does this alter the focus of our questioning or change our impressions of the main insights?
- If anything it reinforces them, especially our interest in the varied performance levels for Germany and China
- It also emphasizes the remarkable consistency of Republic of Korea and France
- At this point, we have definitely established a scent for the story
- We have started to articulate the data questions that best interrogate this data and most likely reflect what the readers of a visualization about this subject will wish to learn

Different Visualization

- We now need a different visual representation
- Using the floating bar we have seen the categorical view of the countries and their performances
- Now, we need to switch our perspective to the other main variable, that of event year, to pursue our curiosities about the transition of medal-winning performances and the transition in ranking of the individual countries across the five Olympic Games
- For this next visual sketch we turn to a line chart
- On this single chart we plot the eight countries, differentiated by color, showing the absolute medal wins from left to right across the five Olympic events

Line Chart



Source: Kirk (2012)

Preparation Line Chart

- Looks a bit messy? This is an exploratory visualization for ourselves
 - You wouldn't and shouldn't publish an isolated, cluttered, and poorly-annotated chart like this to convey a story to others
 - For our exploration, quick and dirty is absolutely fine
- Decision to place all countries onto one graphic is to enable visibility of the interesting transitions, the crossovers, the seemingly cluttered parts, and the empty parts
- You could separate each country out into its own line chart and assess a matrix of eight small-multiples
 - However, this would only show you the individual country stories
 - Our keen interest here is in the relationship between the countries

Insights Line Chart

- The chart shows how Germany's (blue) wide range of results, actually reflects their general decline in medal winning levels and, by extension, their relative rank
- By contrast, China's wide distribution shows a country on the rise over the past four games at least
- The extended fascination of this trend would be whether they will catch up and possibly overtake the US
- Russia can be seen to have moved up and down over the years and has now been overtaken by China
 - Chunk of white space for the 2008 results either side of the Russian value, leaving them quite comfortably in third
- The UK demonstrated a very similar pattern of improvement relative to the Chinese over past five events
- Sometimes no change is as interesting as some change and, in this respect, the consistency of Republic of Korea is quite stark given the different generation of competitors who will have contributed to those totals

Preparing the Story

- Otherwise there is nothing else really of significant interest
- The charts have served their purpose in discovering and confirming some relevant and interesting stories concerning the contrasting experiences of China, Germany and, potentially, the Republic of Korea
- Of course, sometimes you simply may not find a story
 - There just might not be anything of substance to convey to others visually, in which case a table of data may prove to be the most appropriate solution
- However, we have found our stories, so how do we tell them?
- As a bridge the introduction of more visualization types, let's attempt a quick solution

Choosing the Chart Type

Amanda Cox

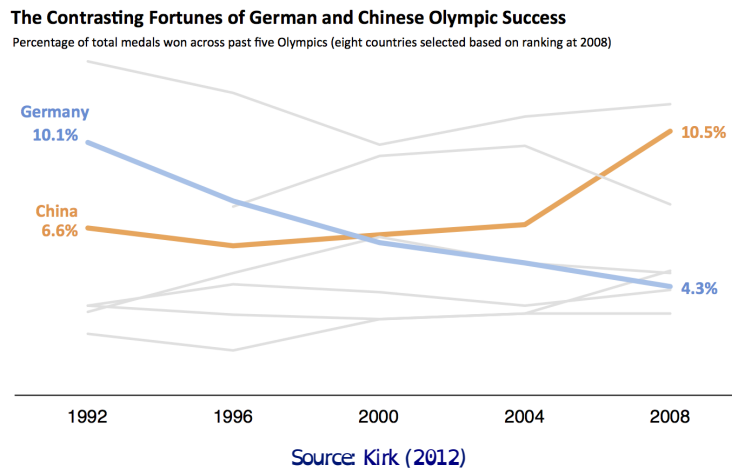
Different forms do better jobs at answering different questions (Source: Kirk (2012))

- Let's reduce the story to a simple contrast between China and Germany
- Our main data question will be something like "how have the medal-winning performances of China and Germany compared over the past five events?"
- Most suitable method for giving form to and answering this question will still be a line chart
- Similar to the one we used for the visual analysis, we are trying to show the relationship between these two countries' respective performance over time

Choosing the Parameters

- The design execution will be different
- This time we're conveying the story to others, so we need to refine the visuals in order to make it an explanatory piece:
 - We need to elevate the important features of the main story and relegate any background context and secondary content
 - We need to ensure that there are annotations for labels, values, and captions so the reader is entirely clear about what is being communicated

Example: Visualization



Design Considerations I

- We have used the calculated data for medals won as a percentage of the total
 - More appropriate for this story as it helps standardize and contextualize the performance across all events in a more comparable way
- Aim here is to provide a clear visual hierarchy emphasizing the two main countries in our story and diminishing the contextualizing six nations into the background.
 - We could have removed the other six countries but, through the use of a subtle shade of gray, we can still see them well enough to get a sense of the overall rankings
 - That is all we need from them - context

Design Considerations II

- Title neatly frames the story, the subheading describes the chart and the data, and the labels help the reader compare the two countries' relative trajectory
- The use of color attempts to help imply the positive improvement (orange = hot = good) of China and the negative decline (blue = cold = bad) of Germany
- Only the bare minimum chart apparatus (the axis line) is included, once again, to allow the main story to come to the fore

4 Tutorial

Assignment 5.4: Lecture Structure

- Per today, there are 5 course meetings left
- Some issues I want to cover
 - Tools: Visualization & Charting

- Tools: Programming Visualizations
- Visualizing...
 - * When
 - * Where
 - * What
 - * With Whom
- Interactivity
- What are your priorities?
- First run of this course. . . lessons being learned?
 - The good
 - The bad
 - The ugly

References

Literatur

- Börner, K. and Polley, D. E. (2014). *Visual Insights – A Practical Guide to Making Sense of Data*. MIT Press, Cambridge, Massachusetts.
- Fry, B. (2008). *Visualizing Data – Exploring and Explaining Data with the Processing Environment*. O'Reilly, Sebastopol, CA.
- 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.
- 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.
- Yau, N. (2013). *Data Points – Visualization that means something*. Wiley.