ESCI upf.
International Business

Data visualization

Course introduction

Marta Coronado Zamora
18 September 2024

Keep in touch

Theory lessons

Marta Coronado Zamora	Jose F. Sánchez
marta.coronado@prof.esci.upf.edu	jose.sanchez@prof.esci.upf.edu
@geneticament	@JFSanchezBioinf
Institut Botànic de Barcelona (CSIC-CMCNB)	Germans Trias i Pujol Research Institute (IGTP)

Practical lessons

Adrià Auladell
adria.auladell@ibe.upf-csic.es
Institut de Biologia Evolutiva (UPF-CSIC)

Course overview

T1 | Introduction. Perception, illusions, inter-individual variability, ranking of visual features and common pitfalls (Marta Coronado Zamora)

Part 1: Tools for data visualization (Marta Coronado Zamora and Adrià Auladell)
T2 | Basic tools for data visualization (`ggplot2`) - P1, P2, P3 (first assignment)
T3 | Dynamic and interactive (`plotly`, `shiny`) - P4, P5 (second assignment)

Part 2: Complex data and dimensionality reduction (Jose F. Sánchez)
T4 | Introduction: Visualization for exploring complex data & Dimensionality reduction - P6
T5 | Principal component analysis - P7
T6 | Non-linear projections: t-SNE - P8 (third assignment)
T7 | Non-linear projections: UMAP - P9 (fourth assignment)

Evaluation

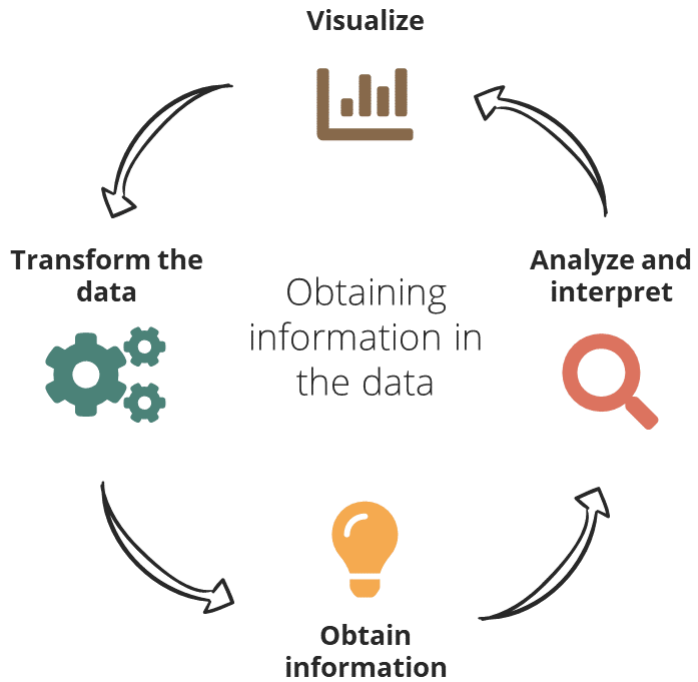
- **10% participation**
  - Tools (Marta and Adrià, 10 sessions) ~5%
    Individual submission at the end of each session
  - Concepts (Jose, 10 sessions) ~5%

- **40% group assignments** (minimum grade 4/10)
  - 4 assignments, each 10%

- **50% final exam** (minimum grade 4/10 in both parts)

Essential principles

- **Data visualization** is the **graphical representation** of **data**
- The main goal is **communicating information clearly** and **effectively**
- Both **aesthetic form** and **functionality** need to go hand in hand

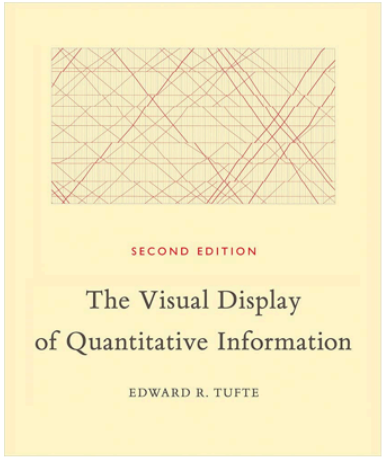


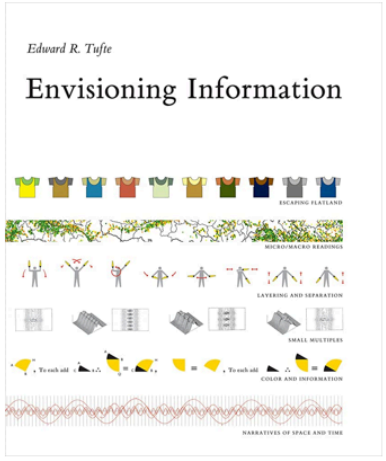
Data insights: a visualization (Aisch 2019)

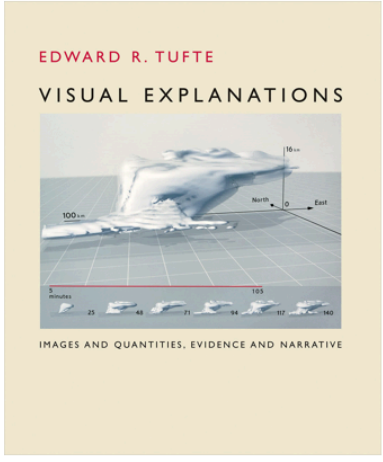
Graphical integrity — Edward Tufte (1942-)

These are Tufte's 6 principles:

1. **Comparisons**: show comparisons to depict contrasts and differences
2. **Causality**: demonstrate how one or more independent variables impact or influence dependent variables
3. **Multivariate**: combine various data
4. **Integration**: incorporate various modes of information
5. **Documentation**: include attribution, detailed titles, and measurements (scales)
6. **Context**: describe the before and after state







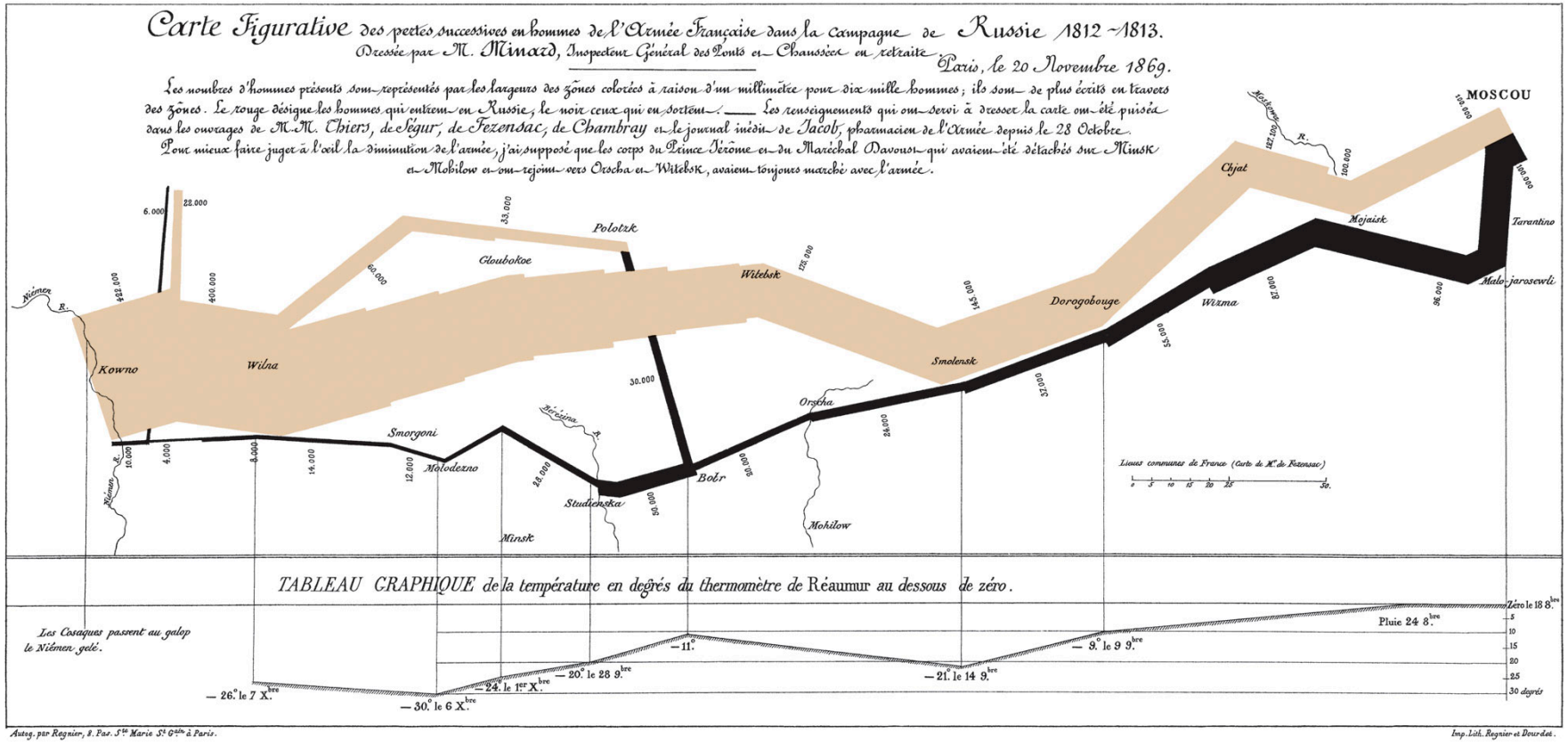
“Above all, show the data”

Chartjunk	Excellence
Excessive and unnecessary use of graphical effects in graphs	Communication of complex ideas with clarity, precision and efficiency

A chart with a considerable amount of junk in it.

❓ What problems do you identify in this figure?

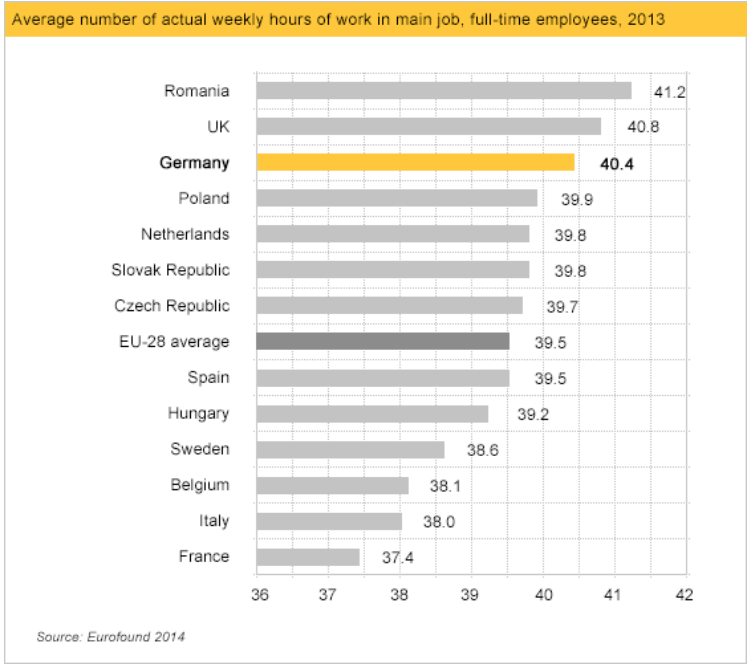
"The best statistical graphic ever drawn"

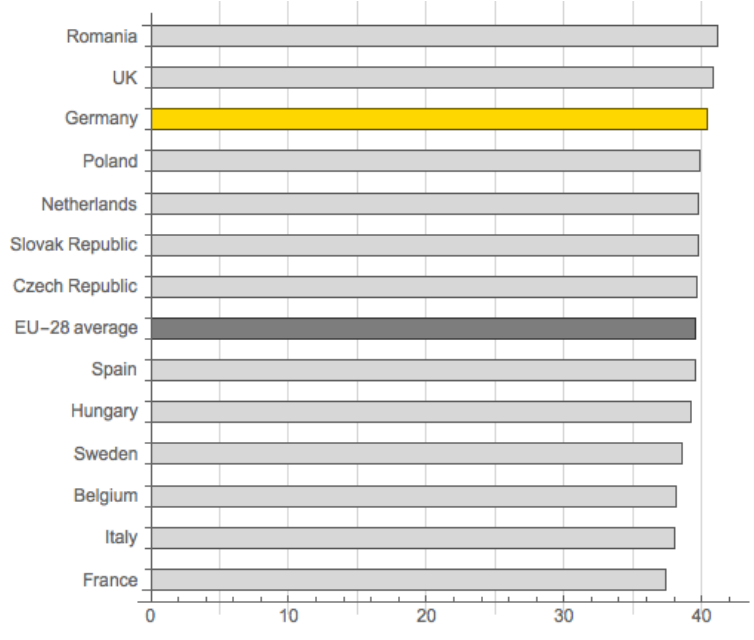


The graphic is notable for its representation in two dimensions of six types of data: the number of Napoleon's troops; distance; temperature; the latitude and longitude; direction of travel; and location relative to specific dates. Wikipedia.

“Above all, show the data”

- Fair axis limits

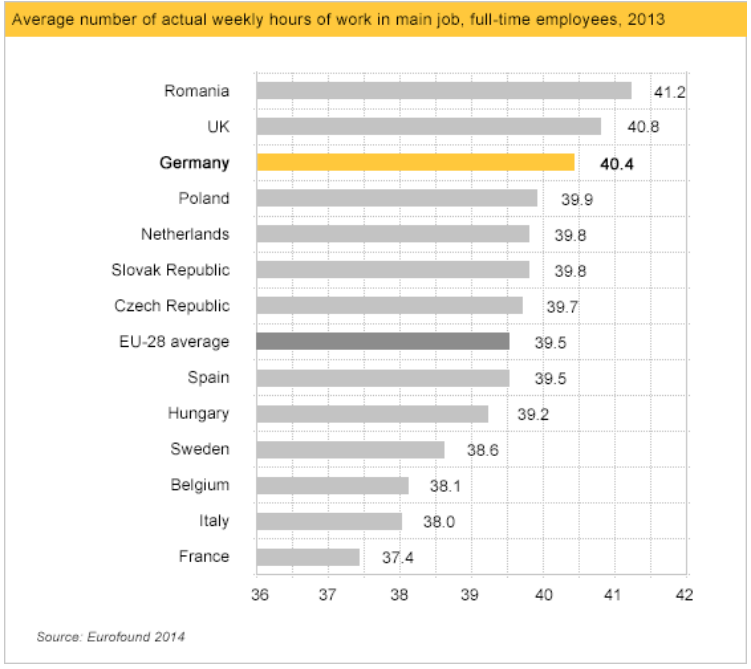




Bar char by the German economic development agency GTAI. German labor market.

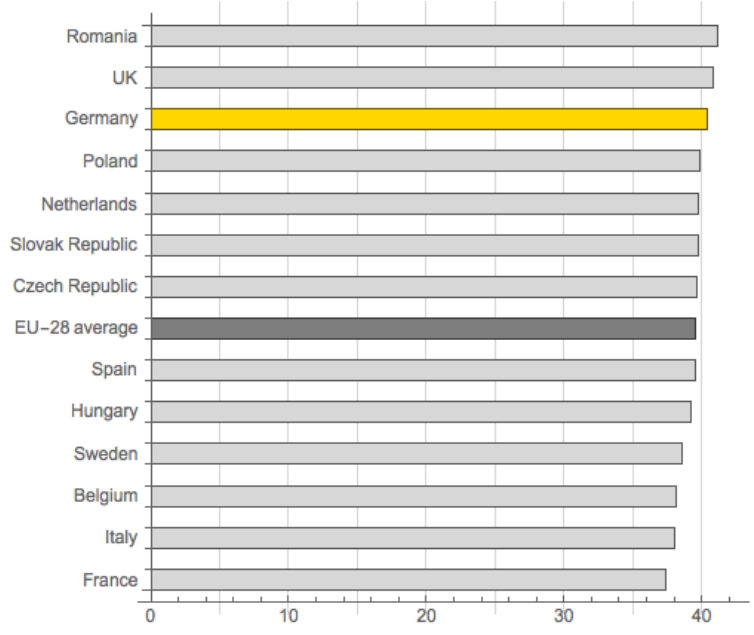
“Above all, show the data”

- Fair axis limits



Bar char by the German economic development agency GTAI. German labor market.

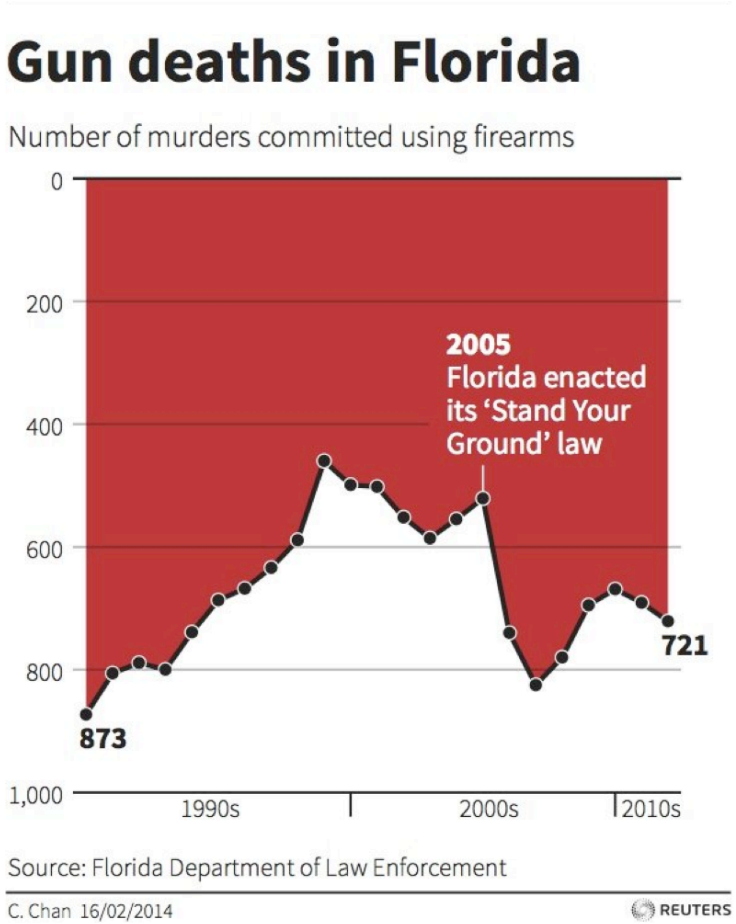
The size of this gap is an illusion.



ℹ Bars in a **bar chart** should (almost) always **extend to zero** (absolute magnitude of values). Not necessary for **line chars** (change in the dependent variable as the independent value changes).

“Above all, show the data”

- Fair axis limits



- Immediate visual impression that gun deaths declined sharply after stand-your-ground legislation was enacted in Florida
- **The decline is an illusion**

“Above all, show the data”

- Fair axis limits
- Encourage the eye to compare values

“Above all, show the data”

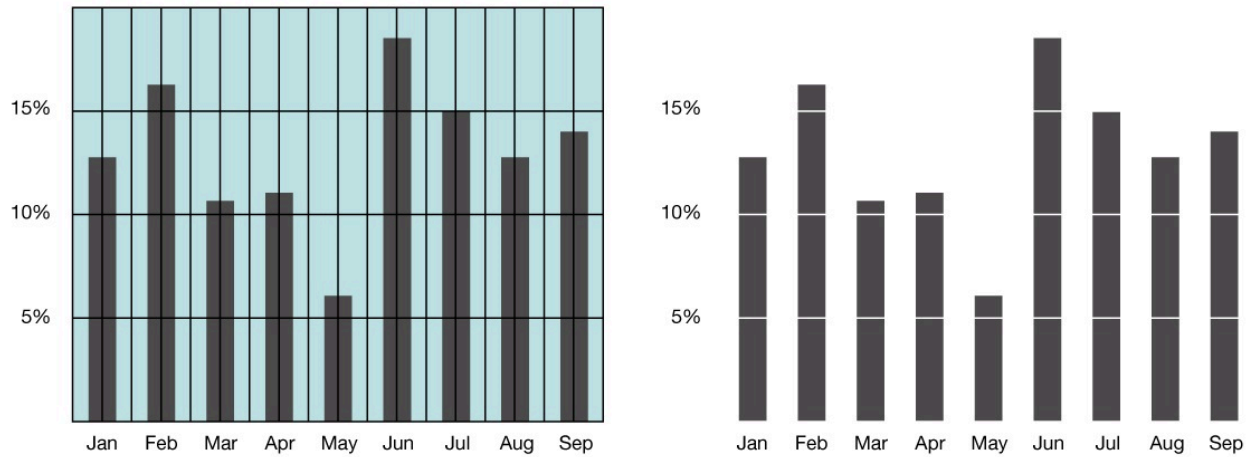
- Fair axis limits
- Encourage the eye to compare values
- Label important parts, axes, units, legends, captions

“Above all, show the data”

- Fair axis limits
- Encourage the eye to compare values
- Label important parts, axes, units, legends, captions
- Consider interactive visualizations (plotly, ggvis, rCharts, ggvis, shiny)

Maximize the data-ink ratio

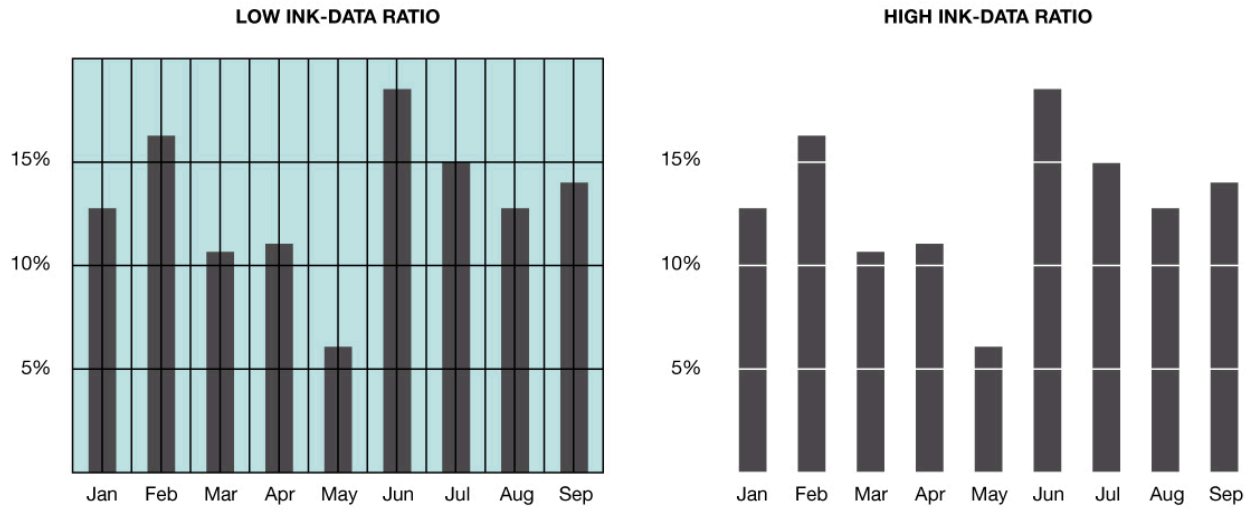
$$Data - ink\ ratio = \frac{Data - ink}{Total\ ink\ in\ graphic}$$



❓ Which graph has higher data-ink ratio?

Maximize the data-ink ratio

$$Data - ink\ ratio = \frac{Data - ink}{Total\ ink\ in\ graphic}$$



❓ Which graph has higher data-ink ratio?

As a general principle, **increase the data-ink ratio to the maximum** that is possible in your graph.

Erase non data-ink

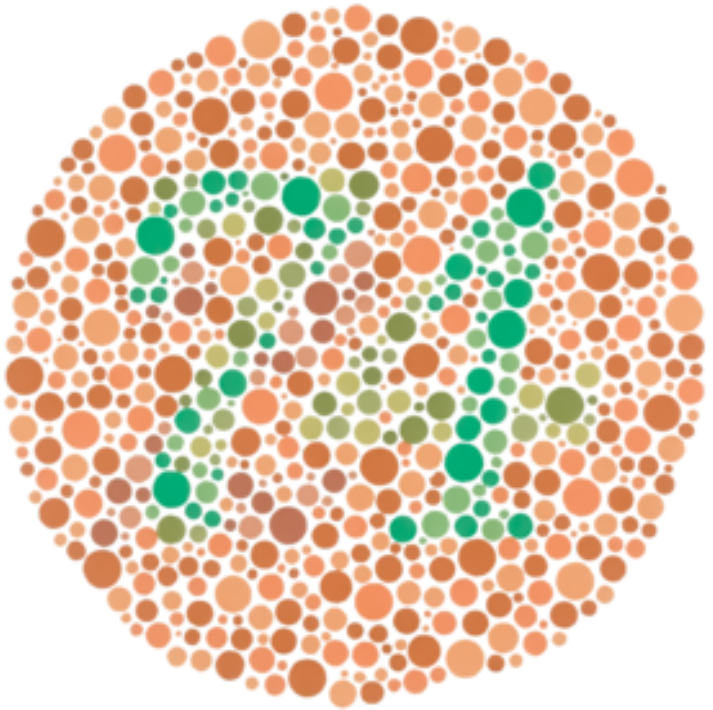
❓ What is the data-ink ratio here?

Design by Nigel Holmes. TIME magazine.

“Nothing is beautiful from every point of view”

Horace

Visual bias #1



❓ **What is the number above?**

Visual bias #2

❓ **Does the dancer spin clockwise or counter-clockwise?**

Visual bias #2

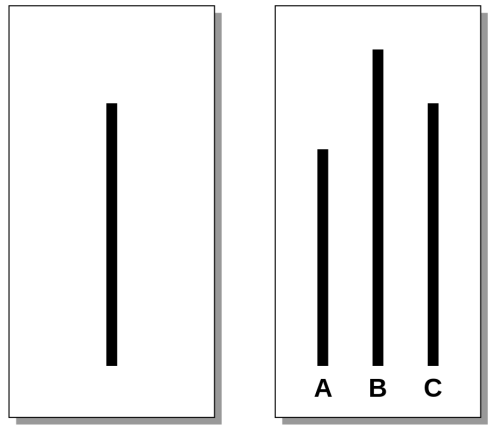
❓ **Does the dancer spin clockwise or counter-clockwise?**

Visual bias #3

❓ Is the dress black and blue, or white and gold?

Visual bias #4

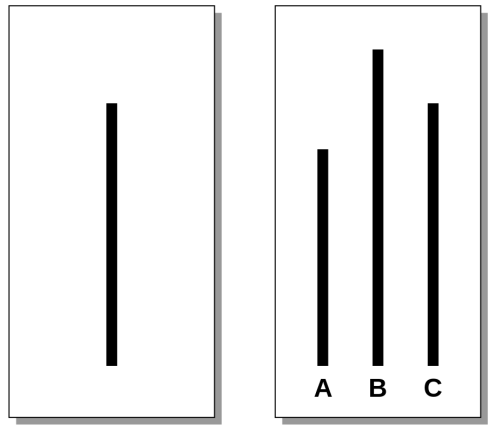
In 1951, the psychologist Solomon Asch asked volunteers to evaluate whether A, B or C has the same length as the bar on the left. In reality, only one person was the subject, the others were actors who were instructed to give the (same) wrong answer.



Visual bias #4

In 1951, the psychologist Solomon Asch asked volunteers to evaluate whether A, B or C has the same length as the bar on the left. In reality, only one person was the subject, the others were actors who were instructed to give the (same) wrong answer.

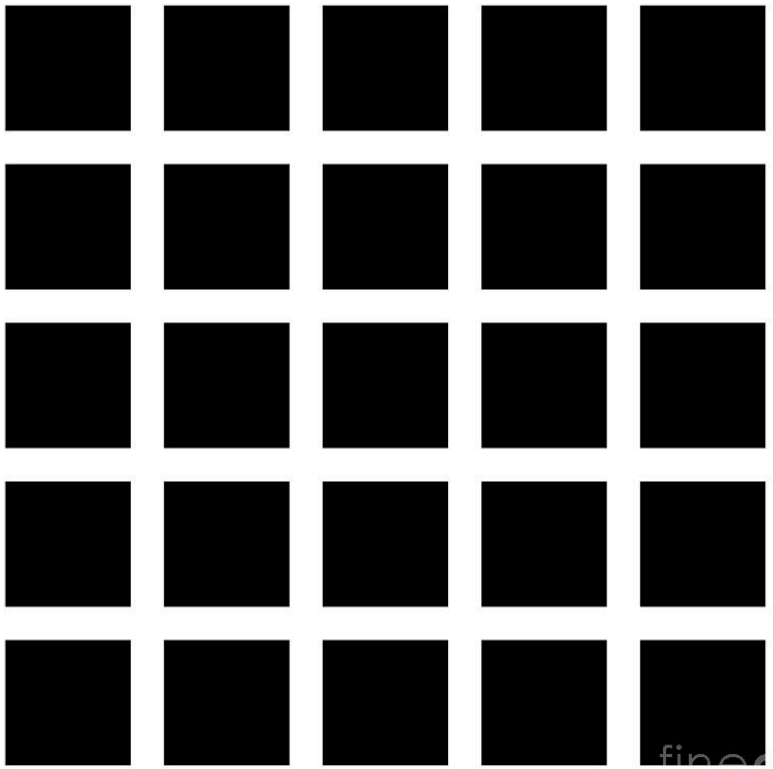
Asch found that 37% of the test subjects also give the wrong answer to conform to the group.



People are different

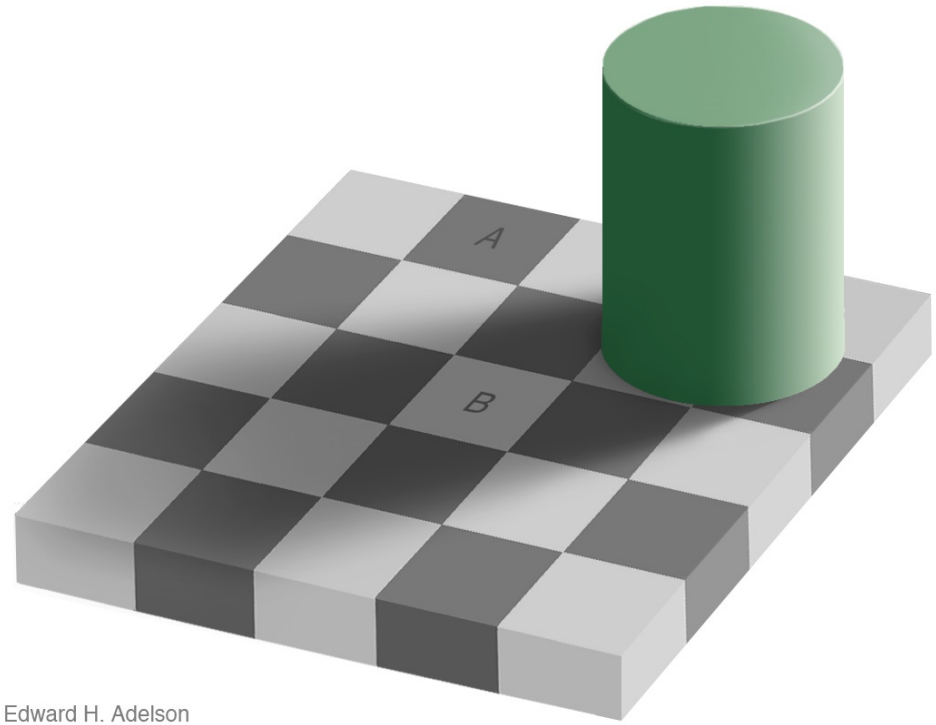
- The brain has some unexpected biases, and the social context matters
- Do not assume people see the same thing as you
- Approximately 8% of men are color-blind

Misperceptions: edges, contrasts and colors



The Hermann grid effect.

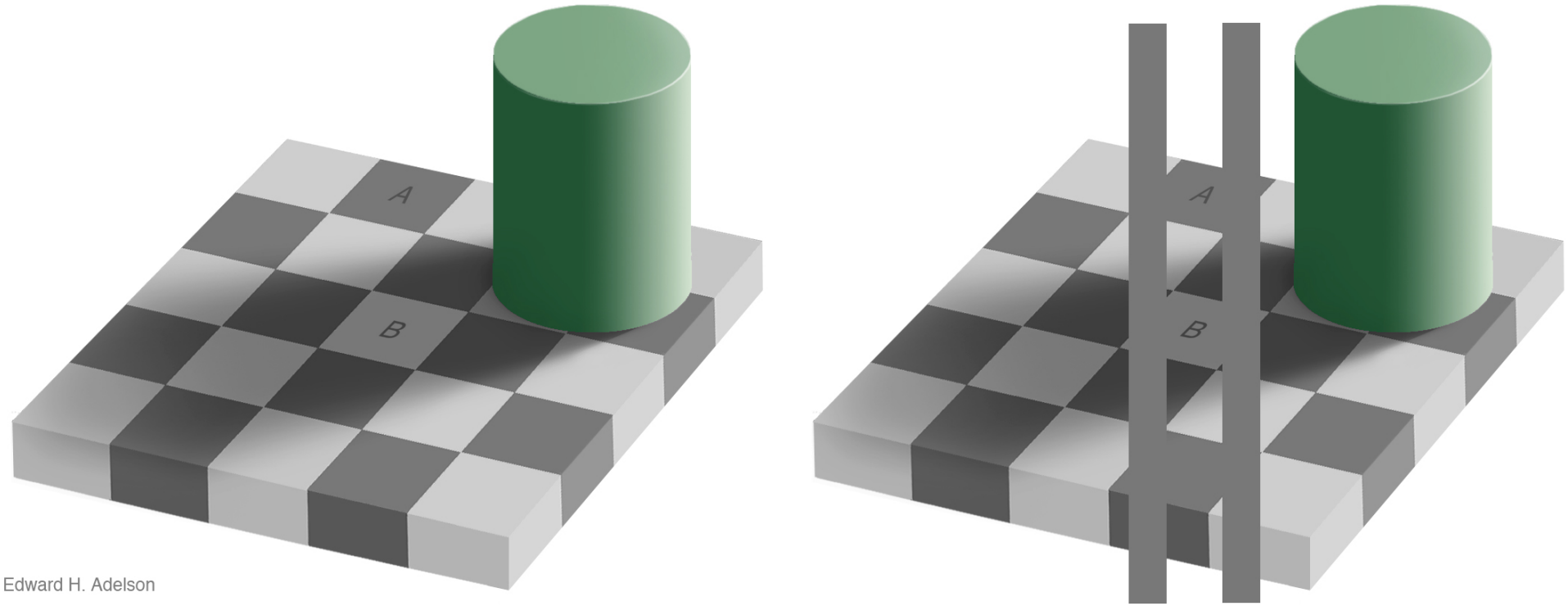
Misperceptions: edges, contrasts and colors



Edward H. Adelson

The checkershadow illusion.

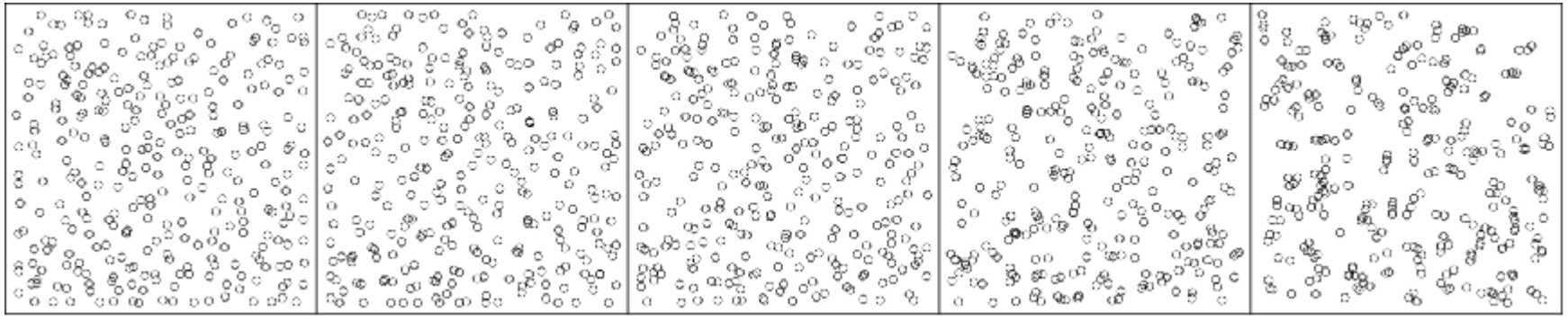
Misperceptions: edges, contrasts and colors



Edward H. Adelson

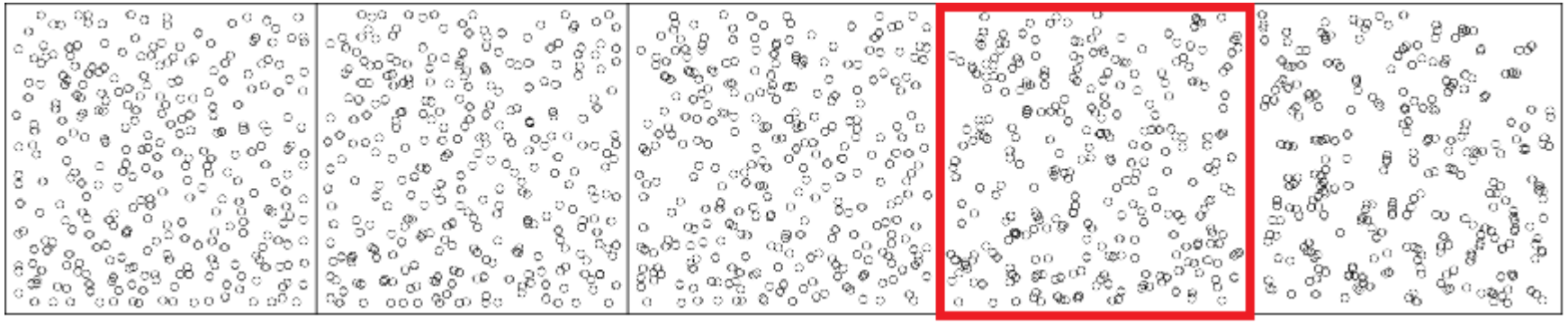
The checkershadow illusion.

Misperceptions: Gestalt rules



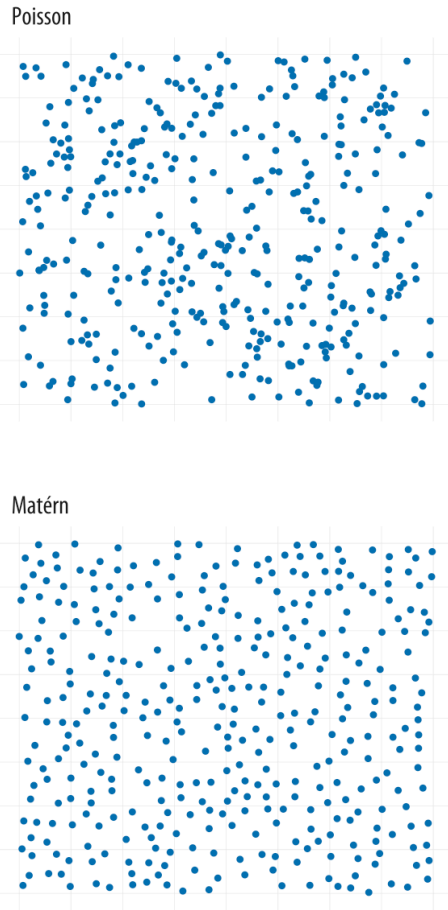
❓ Which distribution is random uniform?

Misperceptions: Gestalt rules



Misperceptions: Gestalt rules

The strong inferences we make about relationships between visual elements from relatively sparse visual information are called “Gestalt rules”.



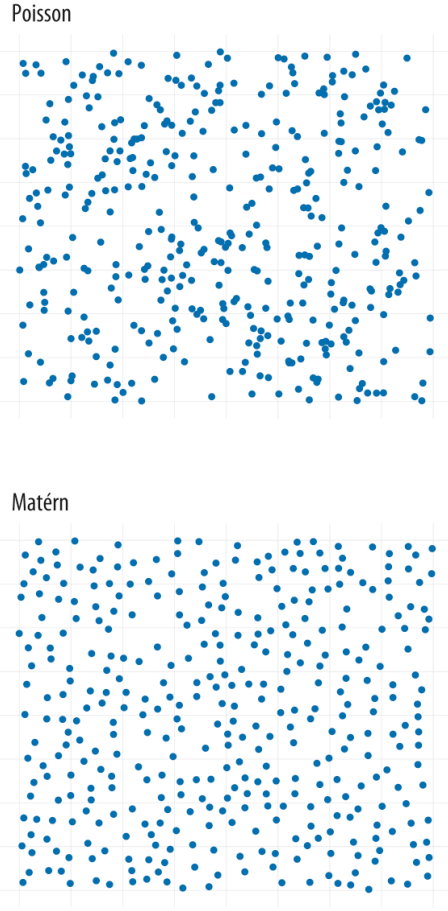
❓ Which distribution is random uniform?

Source: Data Visualization: A Practical Introduction (2018).

Misperceptions: Gestalt rules

The strong inferences we make about relationships between visual elements from relatively sparse visual information are called “Gestalt rules”.

Each panel shows simulated data. The upper panel shows a random point pattern generated by a **Poisson process**. The lower panel is from a **Matérn model**, where new points are randomly placed but cannot be too near already-existing ones. Most people see the Poisson-generated pattern as having more structure, or less ‘randomness’, than the Matérn, whereas the reverse is true!



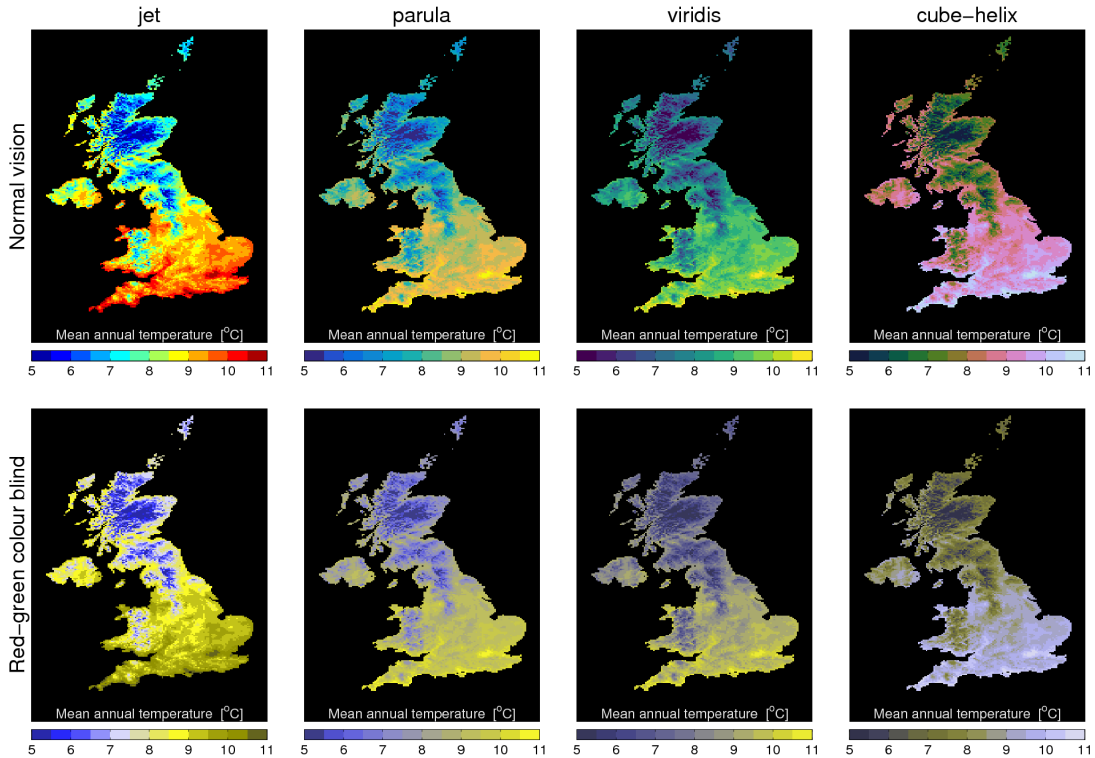
❓ Which distribution is random uniform?

Source: Data Visualization: A Practical Introduction (2018).

The eye is imperfect

- Know the classic biases and avoid them
- Always verify your visual impressions

Standard palettes



UK mean temperature, shown for four different colour scales, for both normal vision (top) and a red-green colour blind simulation (bottom).
The palette “Viridis” is optimized for human vision.

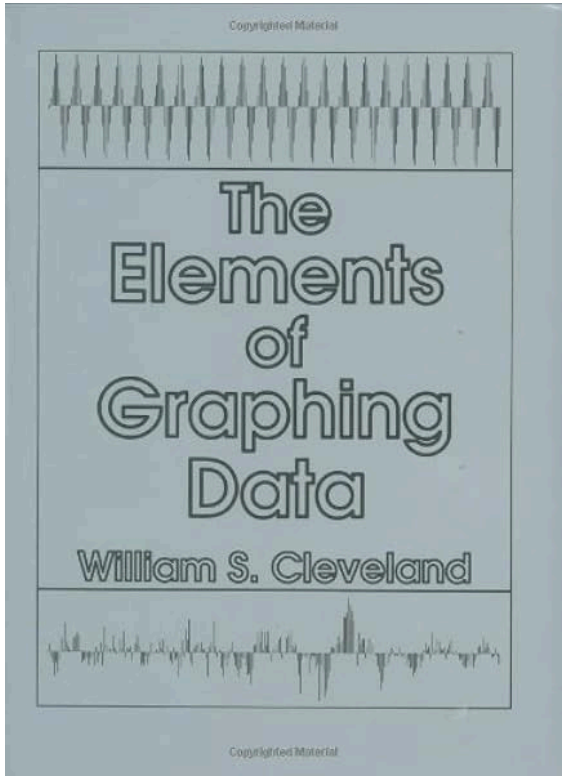
Source: Climate Lab Book.

“I never make the same mistake twice. More like three or four times just to be sure.”

Unknown

Cleveland's three visual operations of pattern perceptions

1. **Detection**: the visual recognition that a geometric object encodes a physical value
2. **Assembly**: grouping of detected graphical elements
3. **Estimation**: visual assessment of the relative magnitude of two or more quantitative physical values



Credit: John Rauser

Levels of estimation

Credit: John Rauser

Three different levels of estimation:

1. **Discrimination**: two values are different `X=Y X≠Y`
2. **Ranking**: something is bigger `X>Y X<Y`
3. **Ratioing**: something is two times bigger `X/Y = ?`

   All of these involves comparison: **efficient comparison between different data points is nearly always the point of a visualization**.

Cleveland's ranking

Credit: John Rauser

Cleveland and McGill (1985) Graphical Perception and Graphical Methods for Analyzing Scientific Data. Science 29(4716):828-833

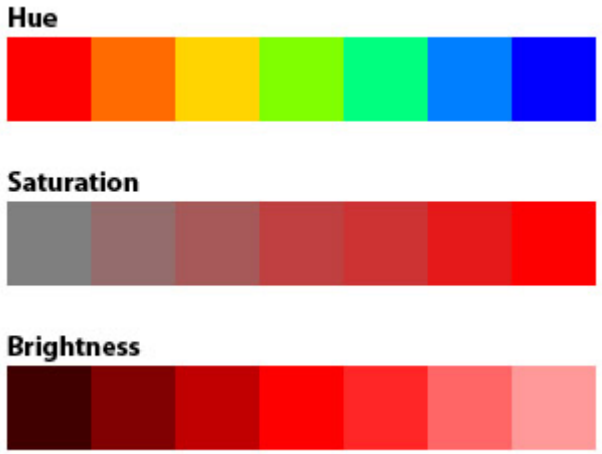
Rank	Aspect judged
1	Position along a common scale
2	Position on identical but unaligned scales
3	Length
4	Angle or slope
5	Area
6	Volume or Density or Color saturation
7	Color hue

How accurate humans are at estimating quantities that are encoded in different ways.

- Seven different ways to encode a quantitative value ranked from most effective to least effective

Color perception

There are three channels that are in encoded in any one color:



Color hue = color

Example with color hue

Credit: John Rauser

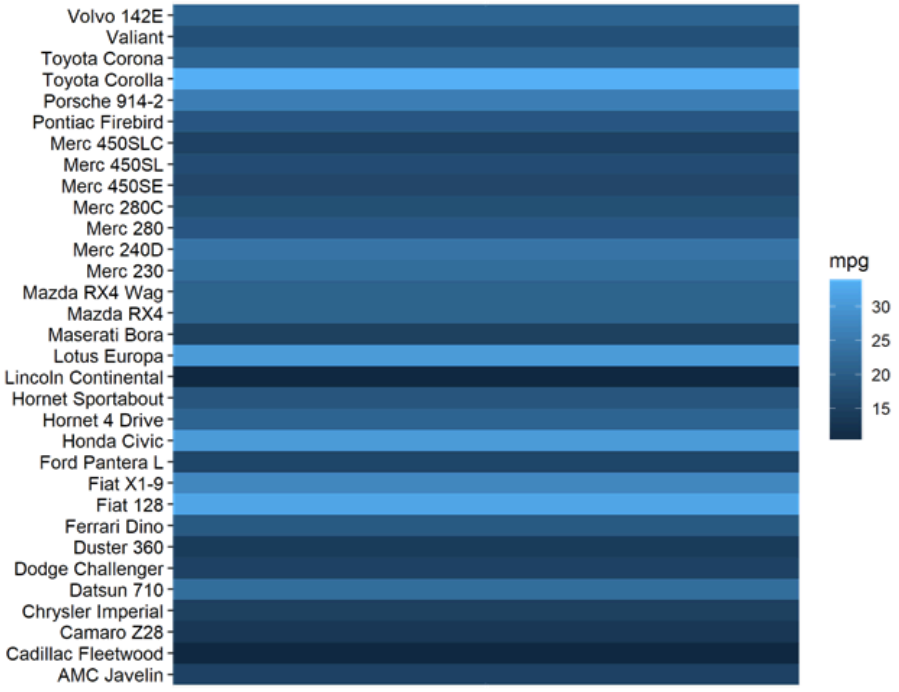


Chart that encodes information using hue - Dataset of 32 cars that were tested in a 1974 issue of Motor Trend magazine. Fuel efficiency in miles per gallon or mpg.

Example with color hue

Credit: John Rauser

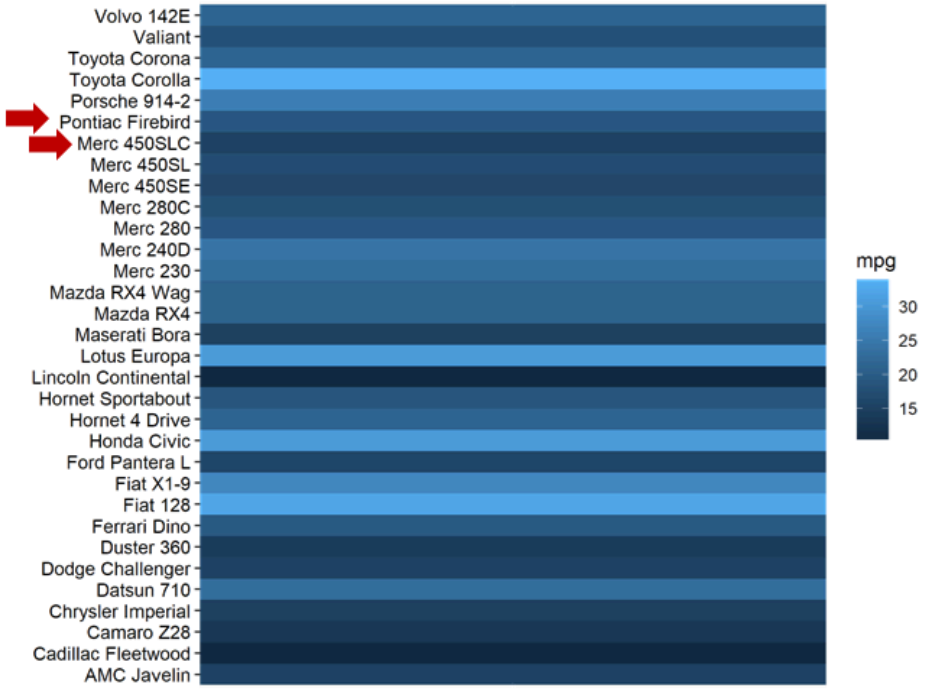


Chart that encodes information using hue - Dataset of 32 cars that were tested in a 1974 issue of Motor Trend magazine. Fuel efficiency in miles per gallon or mpg.

The first Cleveland's estimation task is **discrimination**:
❓ What do you think about these two values: **Pontiac Firebird** vs **Merc450SLC**, are they the same or different?

Example with color hue

Credit: John Rauser

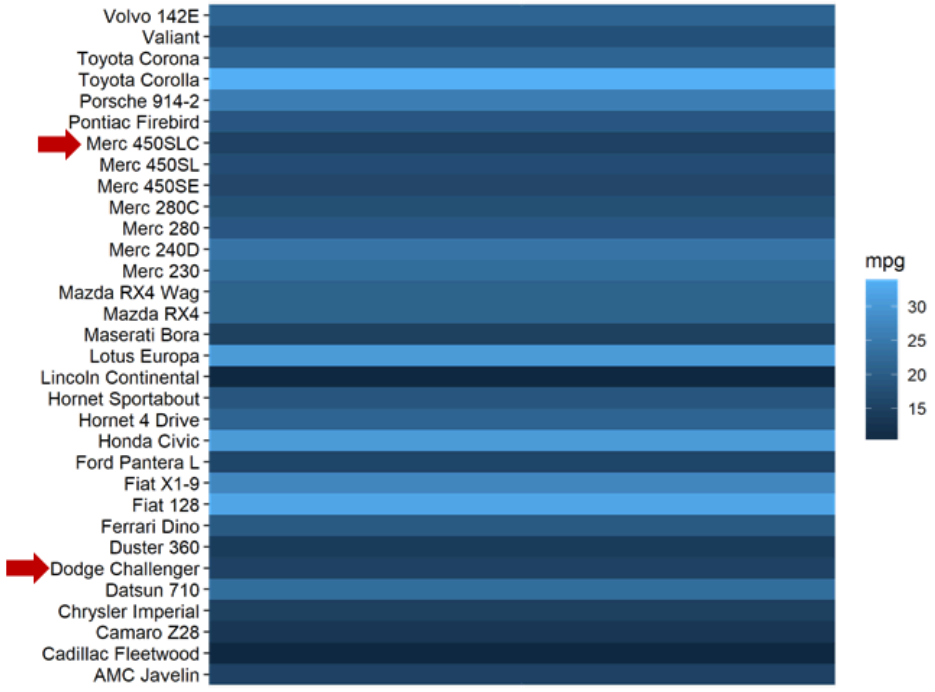


Chart that encodes information using hue - Dataset of 32 cars that were tested in a 1974 issue of Motor Trend magazine. Fuel efficiency in miles per gallon or mpg.

The first Cleveland's estimation task is **discrimination**:
❓ What about **Merc450SLC** vs **Dodge Challenger**, are they the same or different?

Example with color hue

Credit: John Rauser

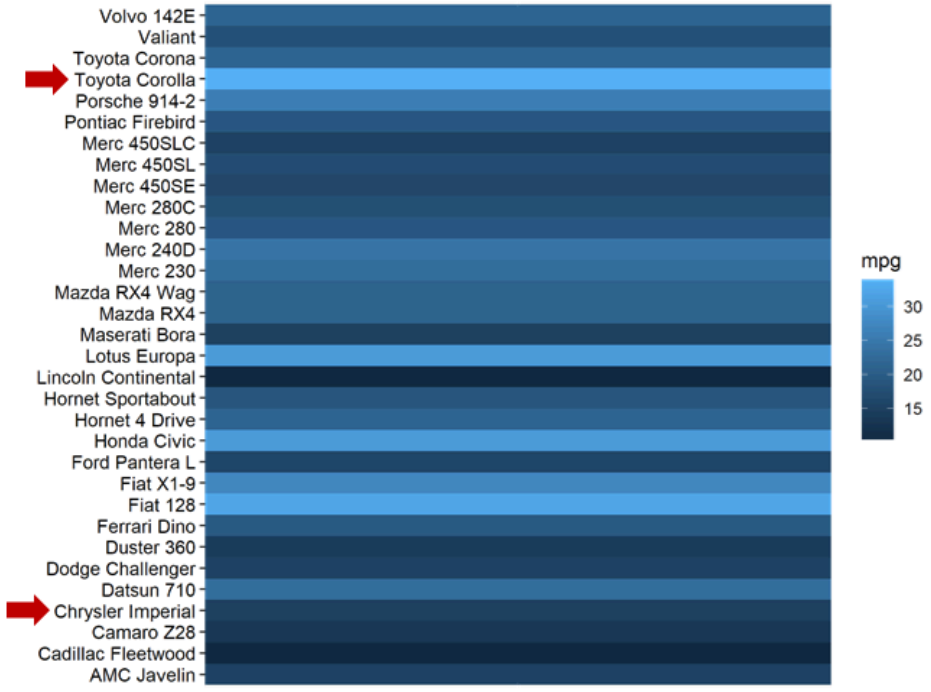


Chart that encodes information using hue - Dataset of 32 cars that were tested in a 1974 issue of Motor Trend magazine. Fuel efficiency in miles per gallon or mpg.

The second Cleveland's estimation task is **ranking**:
❓ What about **Toyota Corolla** vs **Chrysler Imperial**, which has better fuel efficiency?

Example with color hue

Credit: John Rauser

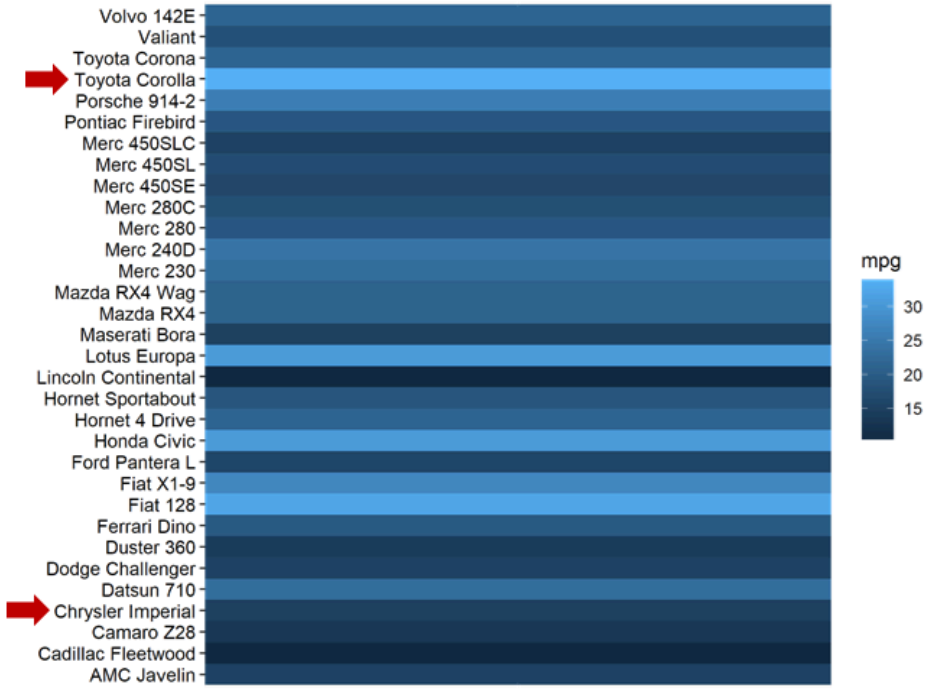


Chart that encodes information using hue - Dataset of 32 cars that were tested in a 1974 issue of Motor Trend magazine. Fuel efficiency in miles per gallon or mpg.

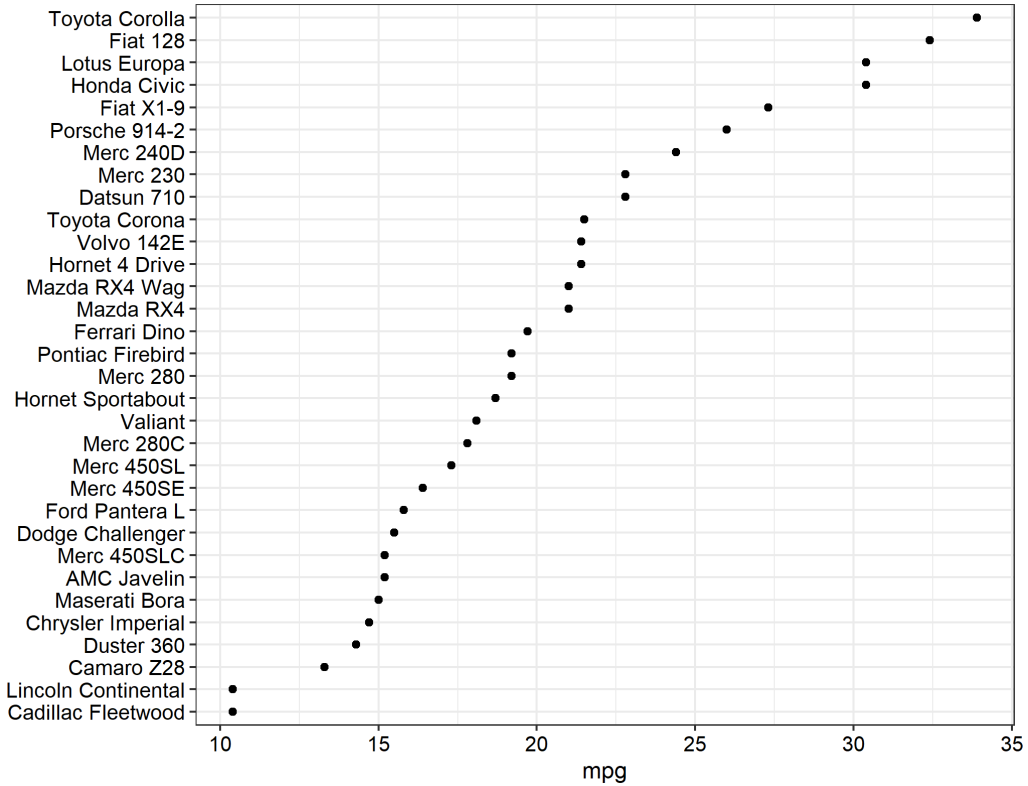
The second Cleveland's estimation task is **ranking**:
❓ What about **Toyota Corolla** vs **Chrysler Imperial**, which has better fuel efficiency?

But you need a **legend** to know if light blue is in the high or low scale: **hue does not have a natural ranking**.

Example with position

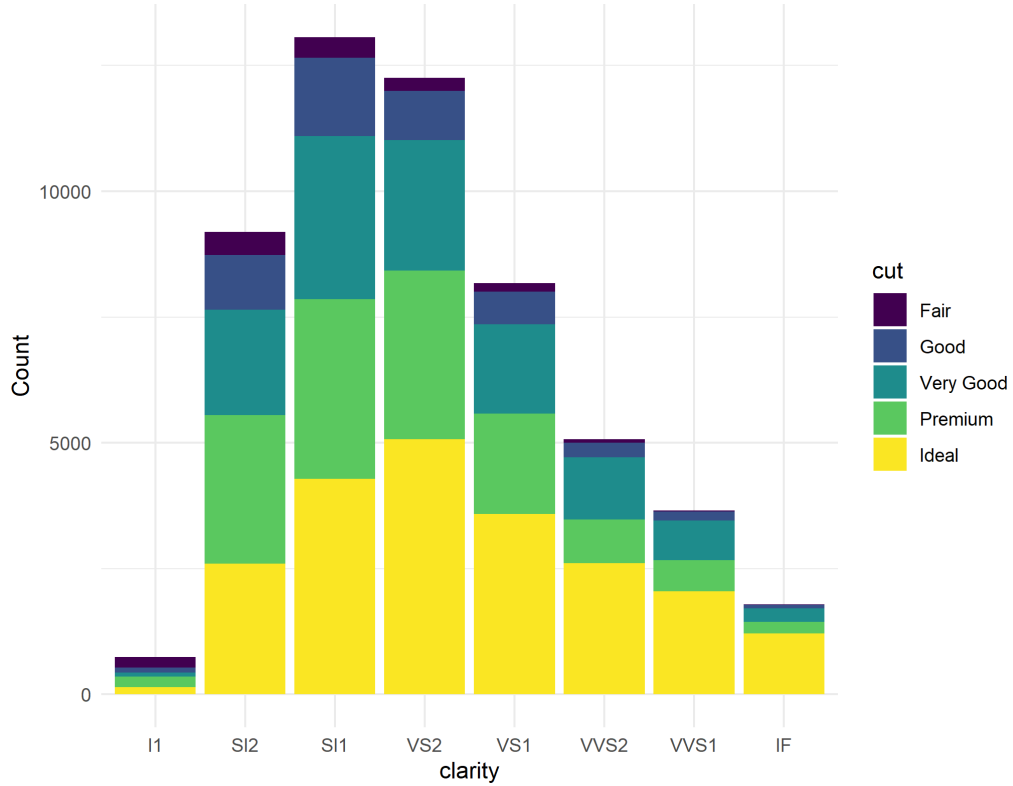
Credit: John Rauser

This is how the data should have been plotted: ranking, discrimination and rationing are trivial!



Why stacking is bad?

Credit: John Rauser

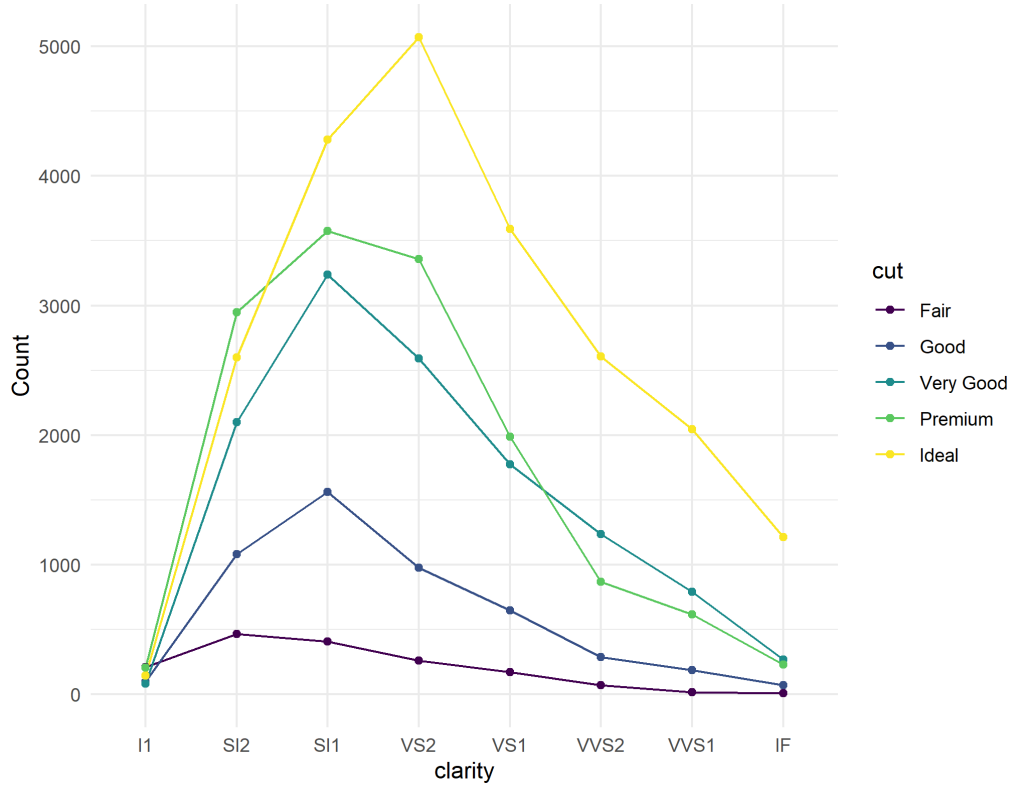


Dataset of 54,000 diamonds. **Stacked bar chart**: Count of diamonds in each combination of cut and clarity.

❓ Are there more SI1 premium cut diamonds or SI2 premium cut diamonds?

Why stacking is bad?

Credit: John Rauser



Dataset of 54,000 diamonds. **Parallel coordinates chart**.

ℹ If you want to communicate the count in each combination of cut and clarity, **encode that information using position on a common scale**.

“A picture is worth a thousand words”

Probably Tess Flanders

The power of images

Mazamet ville morte

In 1973, a journalist realized that the number of **casualties on the road** is the same as the population of the city **Mazamet**.

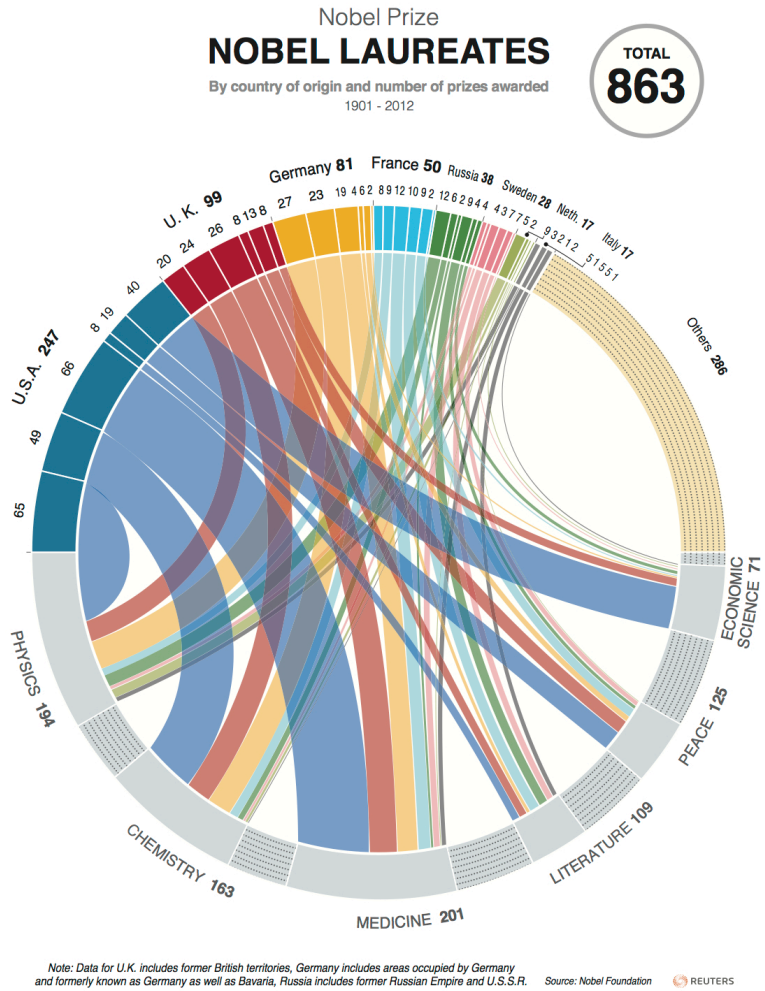
He asked people to play dead, so that people can "***see***" the death burden on the French population.

From 1973, the death toll kept decreasing in France.

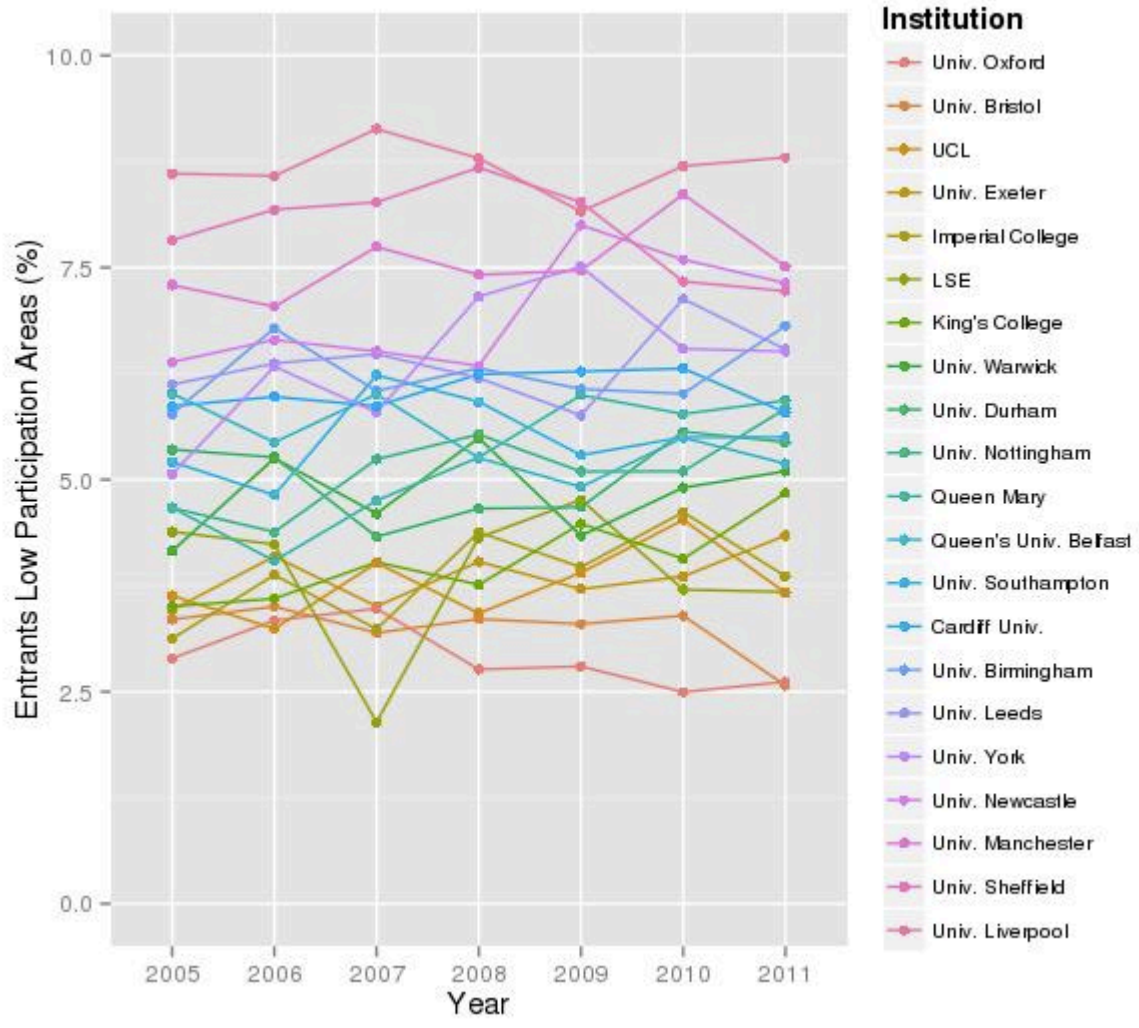
Recap

1. Think about your audience and your message
2. Hide the figure, show the data (or the reverse)
3. See through the eyes of your audience

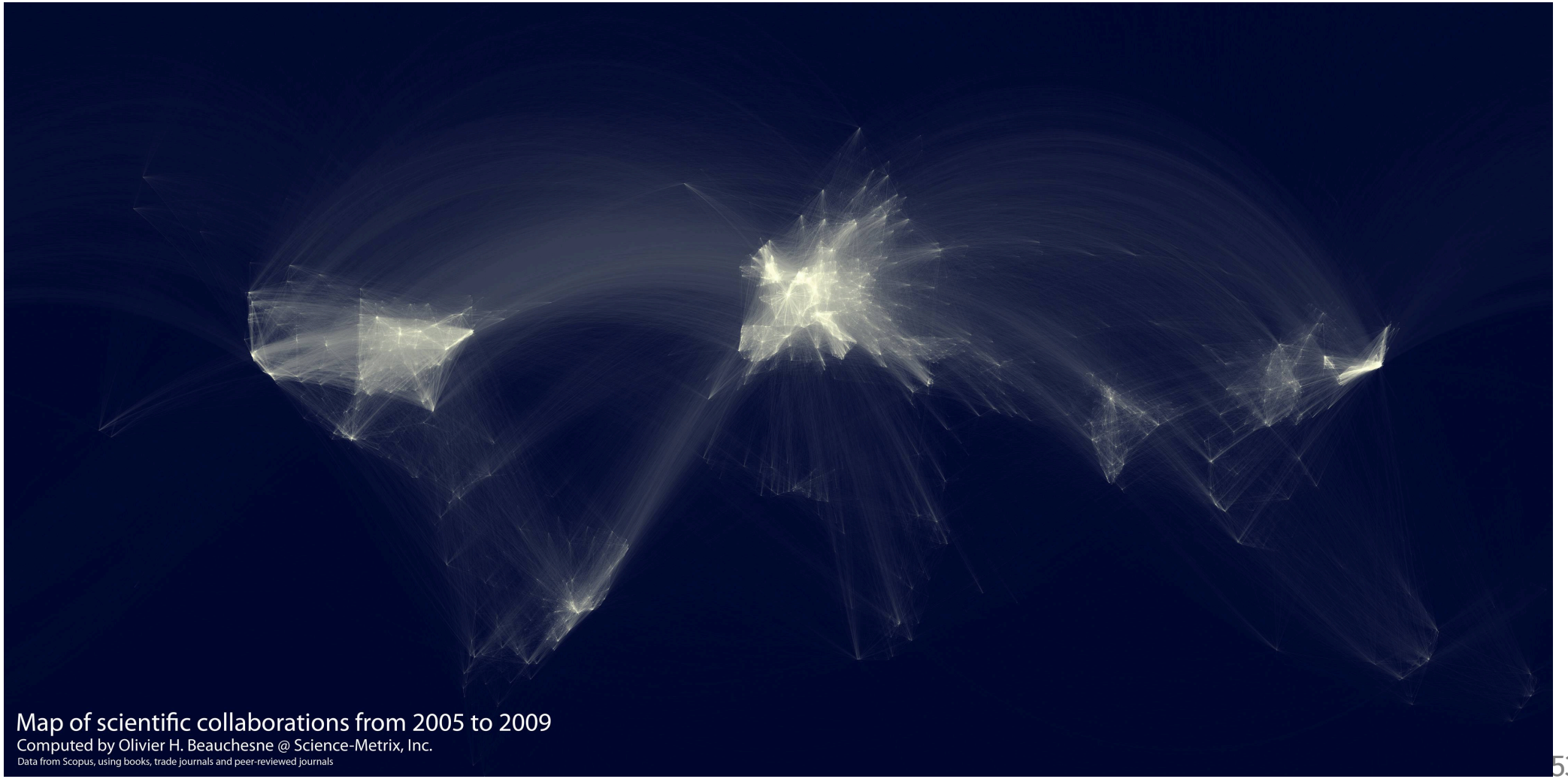
Case study #1


Nobel Prize
NOBEL LAUREATES
By country of origin and number of prizes awarded
1901 - 2012
TOTAL
863
Germany 81
France 50
Russia 38
Sweden 28
Neth. 17
Italy 17
U.K. 99
U.S.A. 247
Others 286
ECONOMIC SCIENCE 71
PEACE 125
LITERATURE 109
MEDICINE 201
CHEMISTRY 163
PHYSICS 194
Note: Data for U.K. includes former British territories, Germany includes areas occupied by Germany and formerly known as Germany as well as Bavaria, Russia includes former Russian Empire and U.S.S.R.
Source: Nobel Foundation
REUTERS

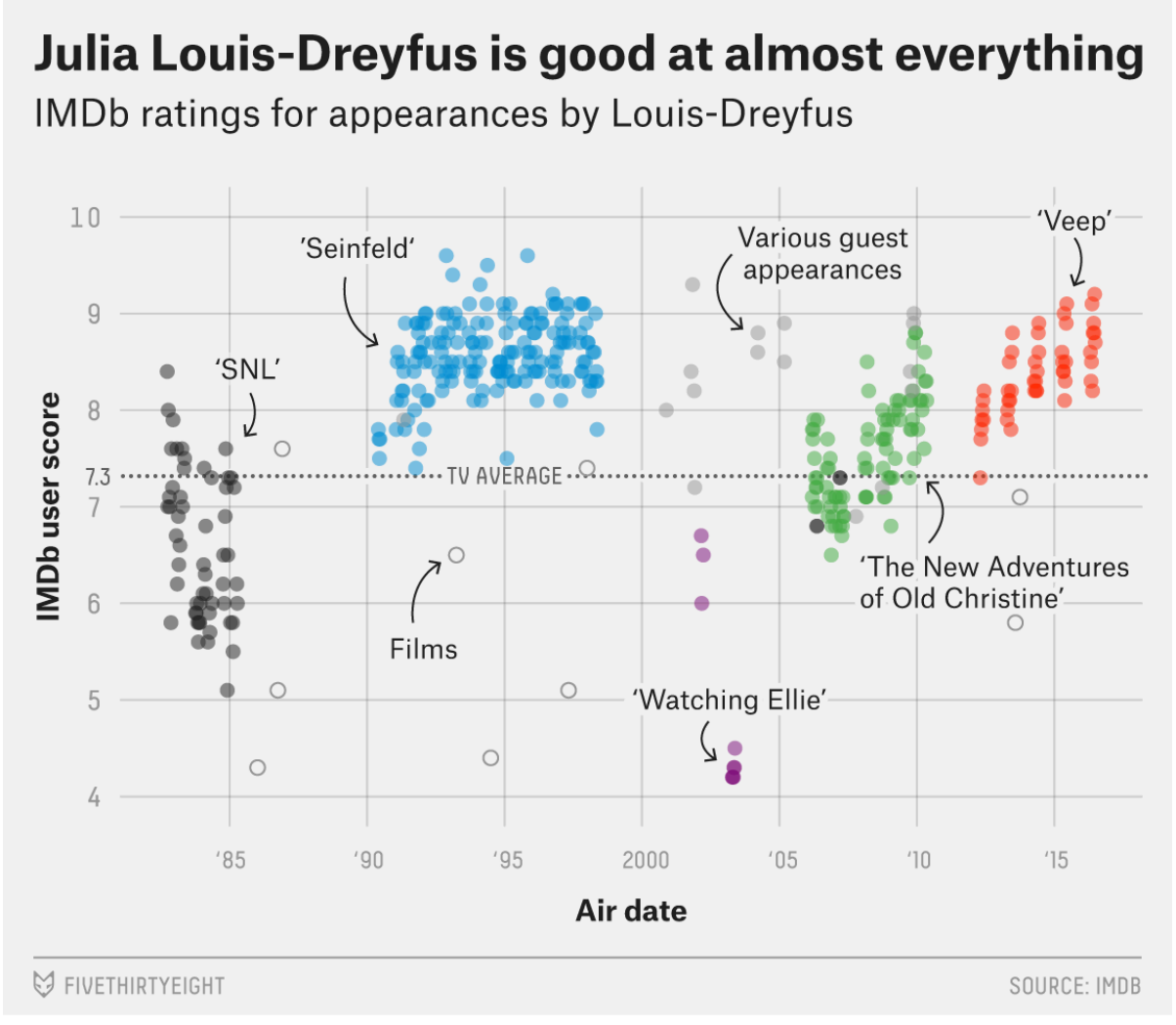
Case study #2


Entrants Low Participation Areas (%)
10.0
7.5
5.0
2.5
0.0
2005
2006
2007
2008
2009
2010
2011
Year
Institution
Univ. Oxford
Univ. Bristol
UCL
Univ. Exeter
Imperial College
LSE
King's College
Univ. Warwick
Univ. Durham
Univ. Nottingham
Queen Mary
Queen's Univ. Belfast
Univ. Southampton
Cardiff Univ.
Univ. Birmingham
Univ. Leeds
Univ. York
Univ. Newcastle
Univ. Manchester
Univ. Sheffield
Univ. Liverpool

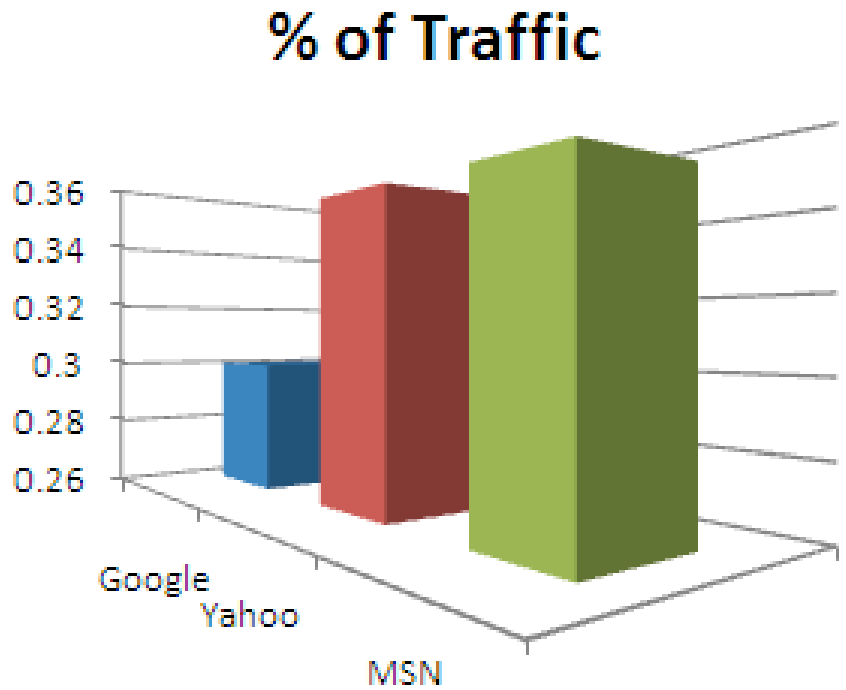
Case study #3


Map of scientific collaborations from 2005 to 2009
Computed by Olivier H. Beauchesne @ Science-Metrix, Inc.
Data from Scopus, using books, trade journals and peer-reviewed journals

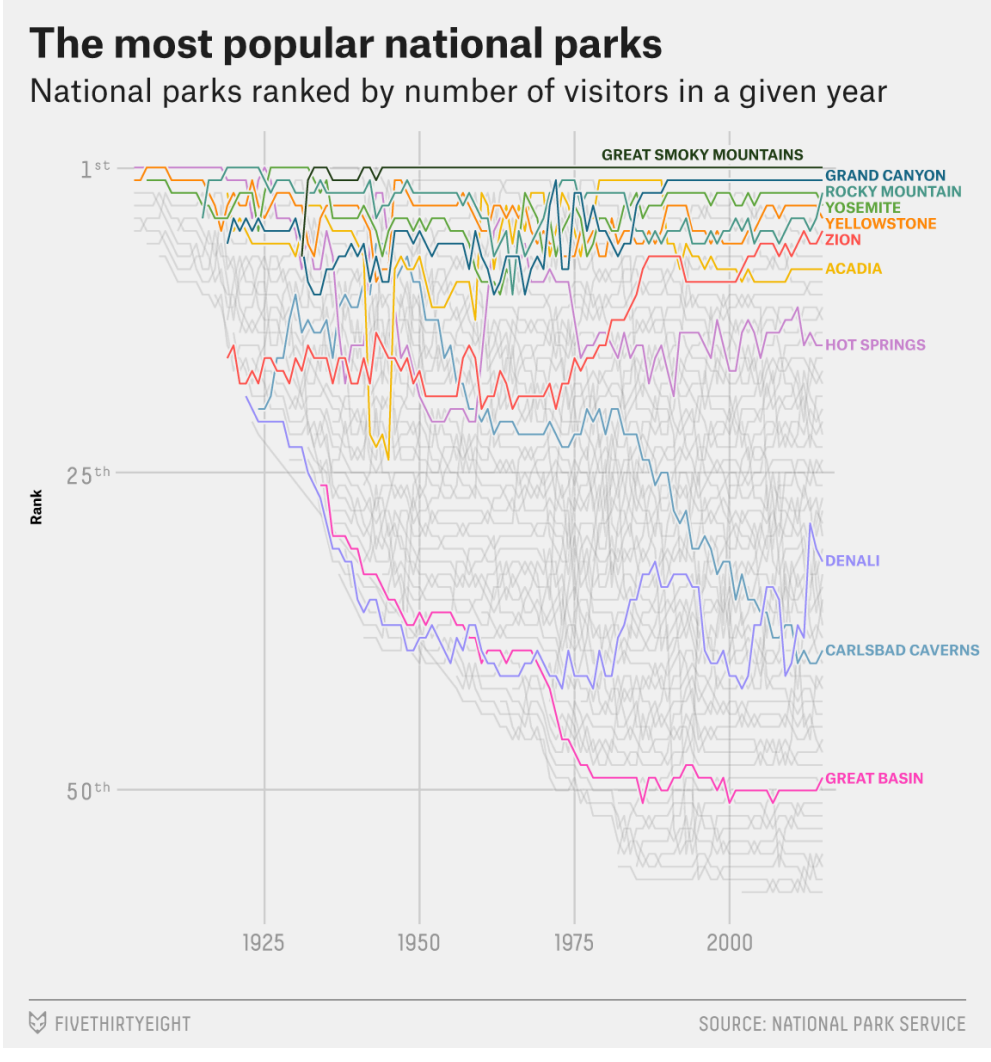
Case study #4


Julia Louis-Dreyfus is good at almost everything
IMDb ratings for appearances by Louis-Dreyfus
IMDb user score
10
9
8
7.3
7
6
5
4
TV AVERAGE
'Seinfeld'
'SNL'
Various guest appearances
'Veep'
'The New Adventures of Old Christine'
Films
'Watching Ellie'
'85
'90
'95
2000
'05
'10
'15
Air date
FIVETHIRTYEIGHT
SOURCE: IMDB

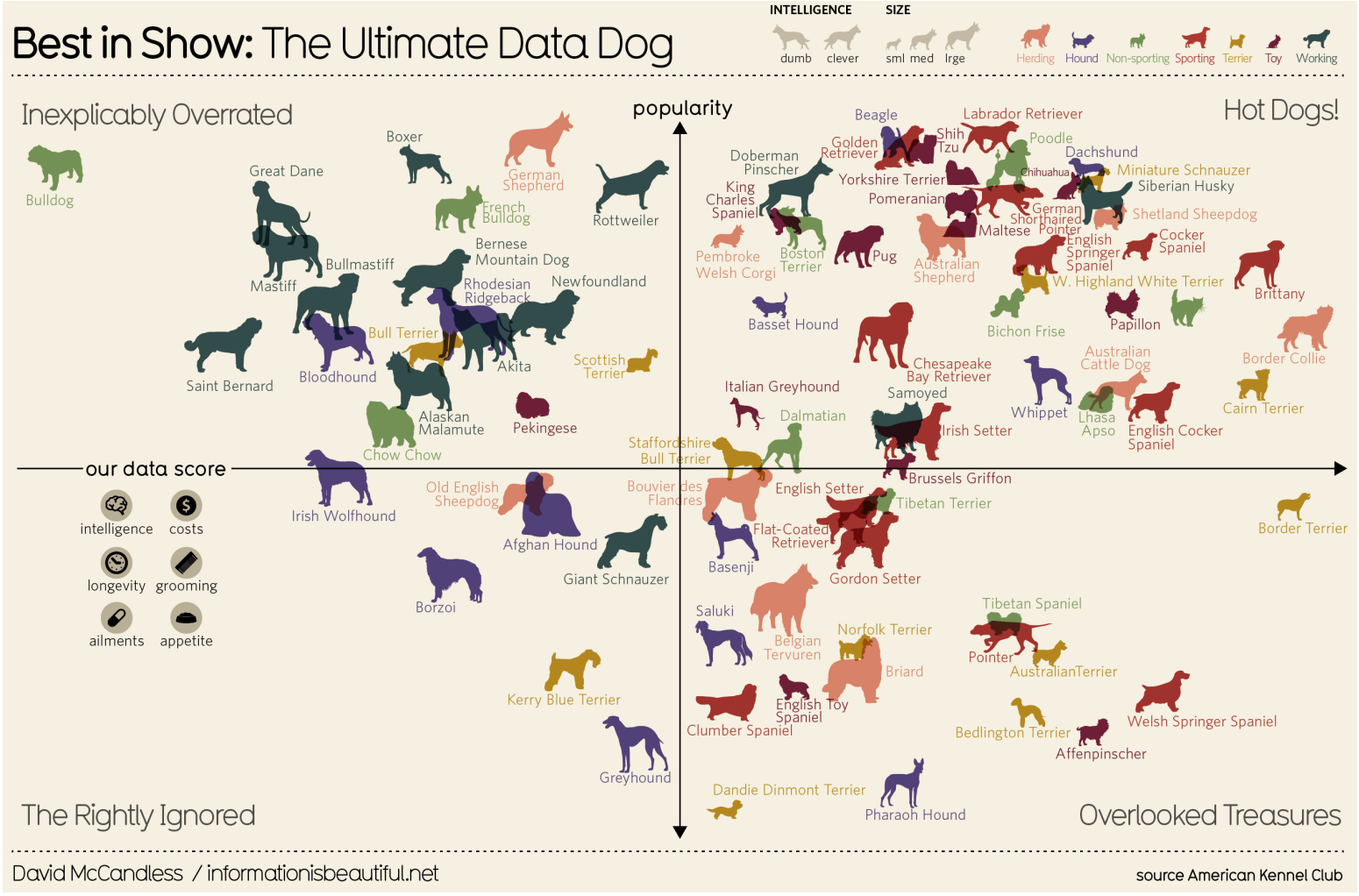
Case study #5



Case study #6


The most popular national parks
National parks ranked by number of visitors in a given year
Rank
1st
25th
50th
1925
1950
1975
2000
GREAT SMOKY MOUNTAINS
GRAND CANYON
ROCKY MOUNTAIN
YOSEMITE
YELLOWSTONE
ZION
ACADIA
HOT SPRINGS
DENALI
CARLSBAD CAVERNS
GREAT BASIN
FIVETHIRTYEIGHT
SOURCE: NATIONAL PARK SERVICE

Case study #7


Best in Show: The Ultimate Data Dog
INTELLIGENCE
dumb
clever
SIZE
sml
med
lrge
Herding
Hound
Non-sporting
Sporting
Terrier
Toy
Working
Inexplicably Overrated
popularity
Hot Dogs!
our data score
intelligence
costs
longevity
grooming
ailments
appetite
The Rightly Ignored
Overlooked Treasures
Bulldog
Boxer
German Shepherd
Great Dane
French Bulldog
Rottweiler
Bernese Mountain Dog
Bullmastiff
Mastiff
Rhodesian Ridgeback
Newfoundland
Bull Terrier
Saint Bernard
Bloodhound
Akita
Scottish Terrier
Alaskan Malamute
Pekingese
Chow Chow
Irish Wolfhound
Old English Sheepdog
Afghan Hound
Giant Schnauzer
Borzoi
Kerry Blue Terrier
Greyhound
Beagle
Labrador Retriever
Golden Retriever
Shih Tzu
Poodle
Doberman Pinscher
Dachshund
Yorkshire Terrier
Miniature Schnauzer
King Charles Spaniel
Pomeranian
Siberian Husky
German Shorthaired Pointer
Shetland Sheepdog
Maltese
Pembroke Welsh Corgi
Boston Terrier
Pug
Australian Shepherd
English Springer Spaniel
Cocker Spaniel
W. Highland White Terrier
Brittany
Basset Hound
Bichon Frise
Papillon
Chesapeake Bay Retriever
Australian Cattle Dog
Border Collie
Italian Greyhound
Samoyed
Whippet
Lhasa Apso
English Cocker Spaniel
Cairn Terrier
Dalmatian
Irish Setter
Staffordshire Bull Terrier
Brussels Griffon
Bouvier des Flandres
English Setter
Tibetan Terrier
Border Terrier
Flat-Coated Retriever
Basenji
Gordon Setter
Saluki
Tibetan Spaniel
Belgian Tervuren
Norfolk Terrier
Pointer
AustralianTerrier
Briard
English Toy Spaniel
Clumber Spaniel
Bedlington Terrier
Welsh Springer Spaniel
Affenpinscher
Dandie Dinmont Terrier
Pharaoh Hound
David McCandless / informationisbeautiful.net
source American Kennel Club

Thanks to Guillaume Fillion for kindly providing his materials.