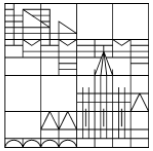
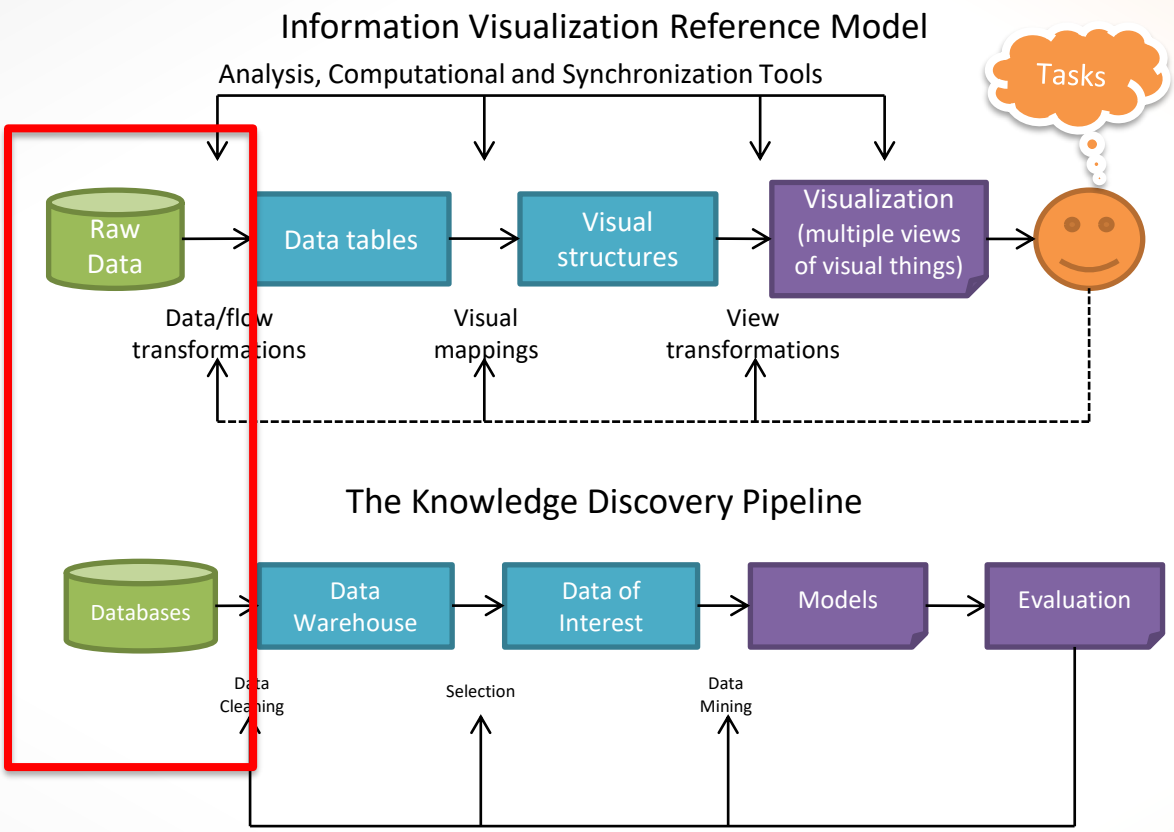
Universität Konstanz



Data Analysis and Visualization

Data Visualization

Data Foundation



It all starts with the data…

Outline

- Types of Data
  - Nominal
  - Ordinal
  - Numeric
- Typical Data Classes
- Data Preprocessing

Outline

- Types of Data
- Typical Data Classes
  - Scalars
  - Multivariate and multidimensional data
  - Vector data
  - Network data
  - Hierarchical data
  - Time-series data
- Data Preprocessing

Outline

- Types of Data
- Typical Data Classes
- Data Preprocessing
  - Data cleaning
  - Normalization
  - Segmentation
  - Data reduction

Nominal

- No quantitative relationship between categories
- Classification without ordering

Ordinal

- Attributes can be rank-ordered
- Distances between values do not have any meaning

Numeric

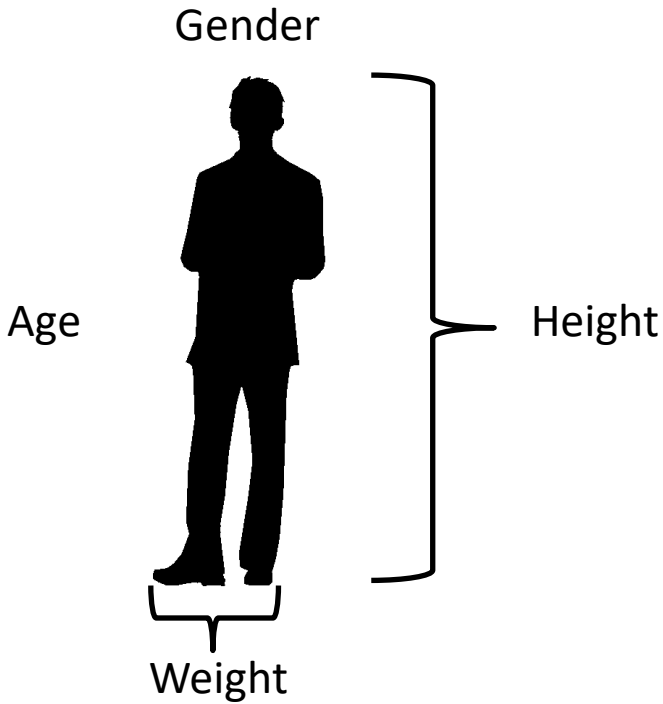
- Attributes can be rank-ordered
- Distances between values have a meaning
- Mathematical operations are possible

Summary: Data Types

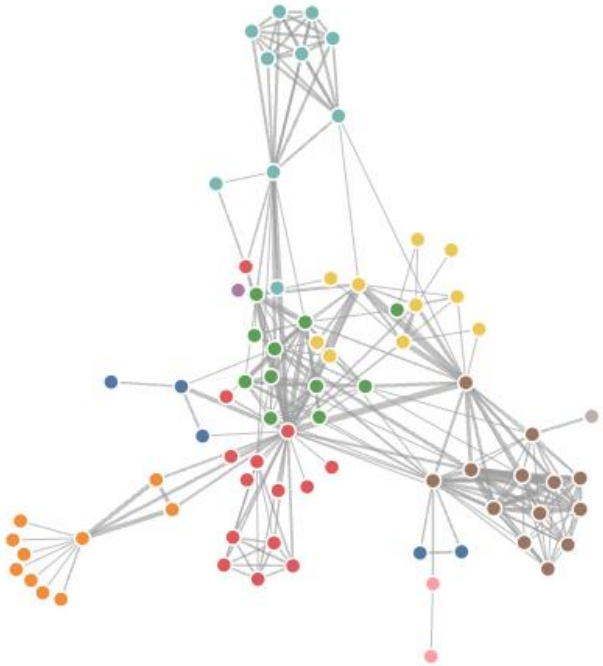
Outline

- Types of Data
- Typical Data Classes
  - Scalars
  - Multivariate and multidimensional data
  - Vector data
  - Network data
  - Hierarchical data
  - Time-series data
- Data Preprocessing

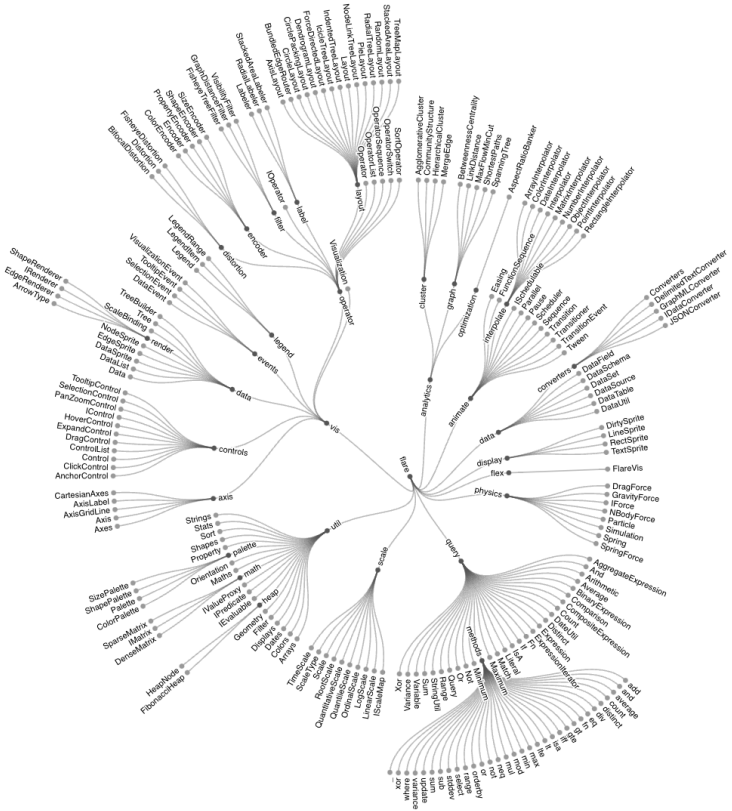
Scalar: *„An individual number in a data record“*



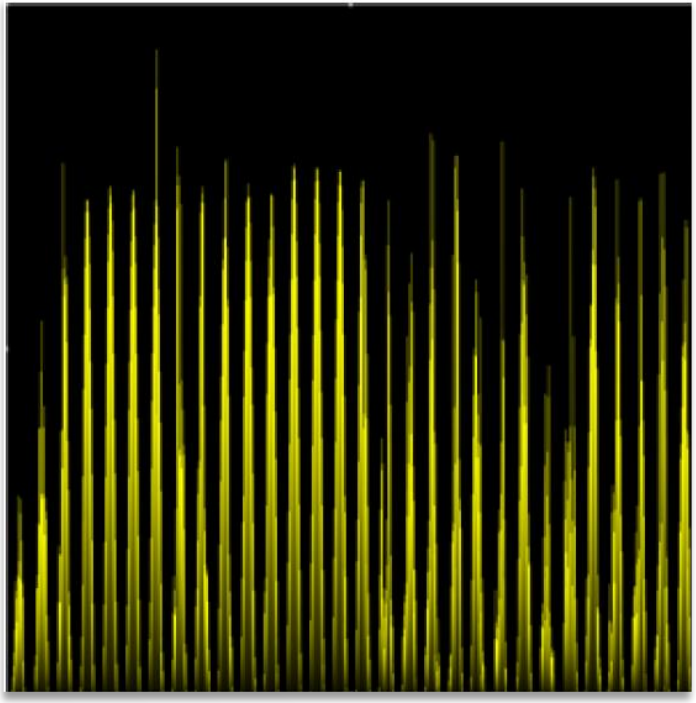
Multivariate Data: „Multiple variables within a single record can represent a composite data item“



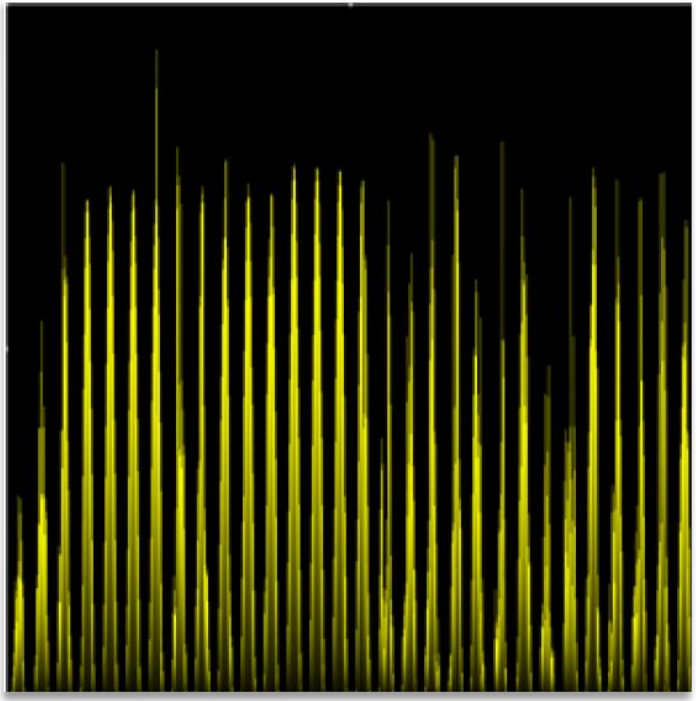
Network Data: *„Vertices on a surface are connected to their neighbors via edges.“*

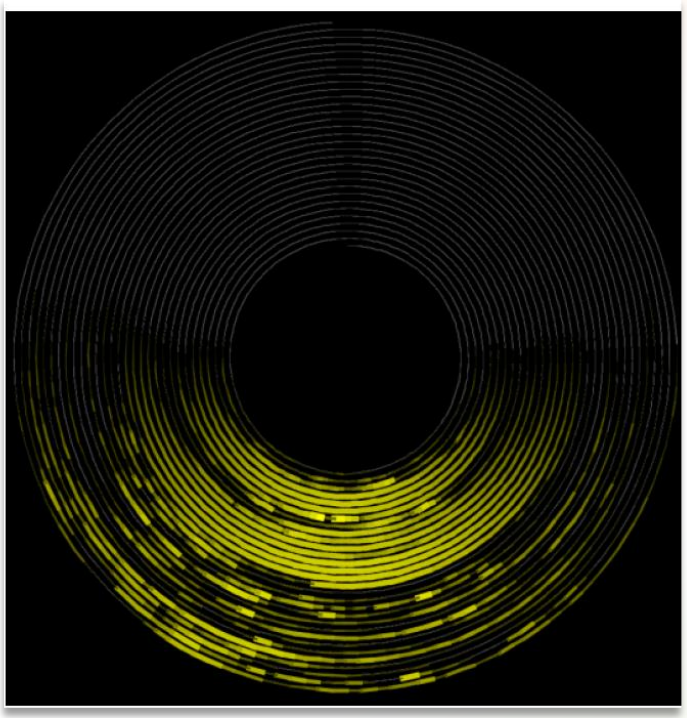


Hierarchical Data: *„[...] relationships between nodes in a hierarchy can be specified by links.“*



Time-Series Data: *„Time perhaps has the widest range of possible values...“*



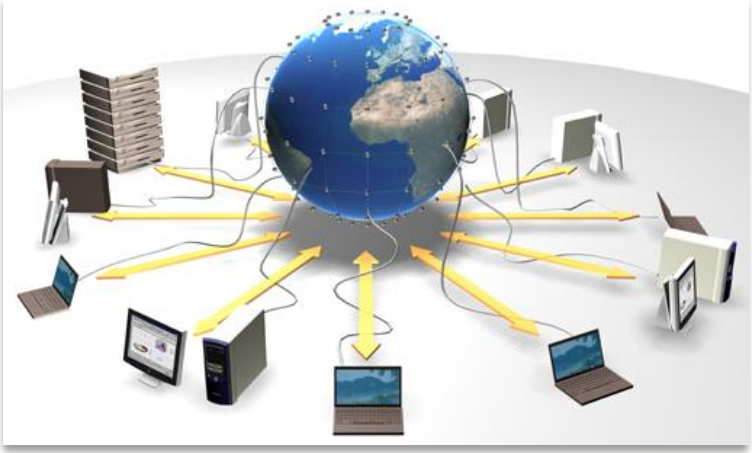


Time-Series Data: *„Time perhaps has the widest range of possible values...“*

Hands-On-Session: Which data class(es) would you use to describe the movement of a flock of ducks?

Outline

- Types of Data
- Typical Data Classes
- Data Preprocessing
  - Data cleaning
  - Normalization
  - Segmentation
  - Data reduction



„Low-quality data will lead to low-quality mining results.“

Data Cleaning – Missing Values

Missing values can have various reasons.

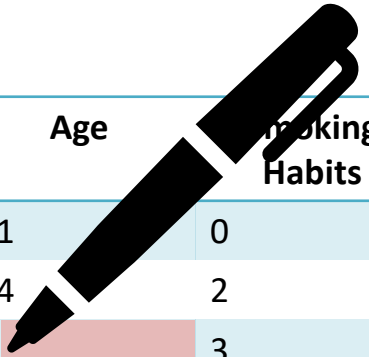
- Someone forgot to fill out a field in a questionnaire like gender or age
- There was a power outage while collecting data with a sensor
- Buffer was full and couldn‘t store more data
- Etc.

Unique ID	Gender	Age	Smoking Habits	Time
323	0	21	0	1h 40s
435	1	34	2	1h 55s
123	0		3	1h 23s
352	1	25	0	1h 43s
674	1	25	1	1h 10s
865	0	18	3	1h 50s
341	1	41	2	1h 33s

Dealing with missing values: what are our options?

Unique ID	Gender	Age	Smoking Habits	Time
323	0	21	0	1h 40s
435	1	34	2	1h 55s
~~123~~	~~0~~		~~3~~	~~1h 23s~~
352	1	25	0	1h 43s
674	1	25	1	1h 10s
865	0	18	3	1h 50s
341	1	41	2	1h 33s

Dealing with missing values: ignore the tuple (delete the record)



Unique ID	Gender	Age	Smoking Habits	Time
323	0	21	0	1h 40s
435	1	34	2	1h 55s
123	0	2	3	1h 23s
352	1	25	0	1h 43s
674	1	25	1	1h 10s
865	0	18	3	1h 50s
341	1	41	2	1h 33s

Dealing with missing values: fill in the missing value manually

Unique ID	Gender	Age	Smoking Habits	Time
323	0	21	0	1h 40s
435	1	34	2	1h 55s
123	0	**-1**	3	1h 23s
352	1	25	0	1h 43s
674	1	25	1	1h 10s
865	0	18	3	1h 50s
341	1	41	2	1h 33s

Dealing with missing values: use a global constant

Unique ID	Gender	Age	Smoking Habits	Time
323	0	21	0	1h 40s
435	1	34	2	1h 55s
123	0	**27**	3	1h 23s
352	1	25	0	1h 43s
674	1	25	1	1h 10s
865	0	18	3	1h 50s
341	1	41	2	1h 33s

Dealing with missing values: what are our options?

Unique ID	Gender	Age	Smoking Habits	Time
323	0	21	0	1h 40s
435	1	34	2	1h 55s
123	0		3	1h 23s
352	1	25	0	1h 43s
674	1	25	1	1h 10s
865	0	18	3	1h 50s
341	1	41	2	1h 33s

Dealing with missing values: what are our options?

Enter value manually

+ For small datasets effective
- Need to know the value
- Time consuming
- Not feasible with large datasets

Use attribute mean

+ Simple to implement
- Not the most accurate approximation of the value

Ignore the tuple

+ Can be easily done
+ No computational effort
- Loss of information
- Unnecessary if the attribute is not needed

Use a global constant

+ Can be easily done
+ Perhaps interesting to know the missing values

Your algorithm should not use this value for further calculations (be careful with "-1")

Use most probable value

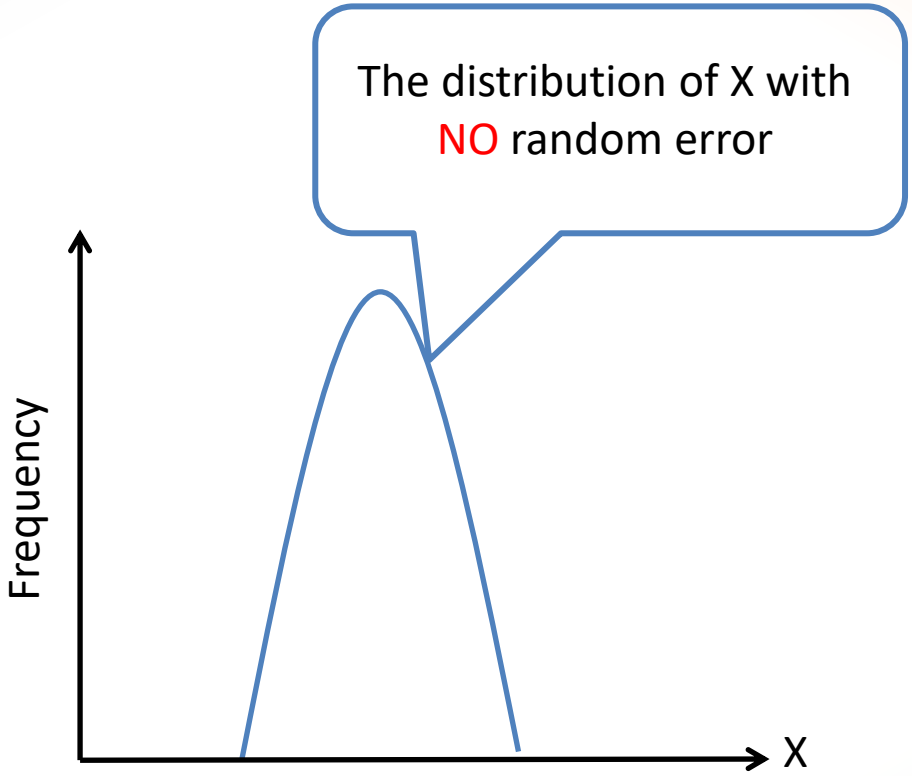
+ Most accurate approximation of the value
- Most computational effort

Dealing with missing values - Summary

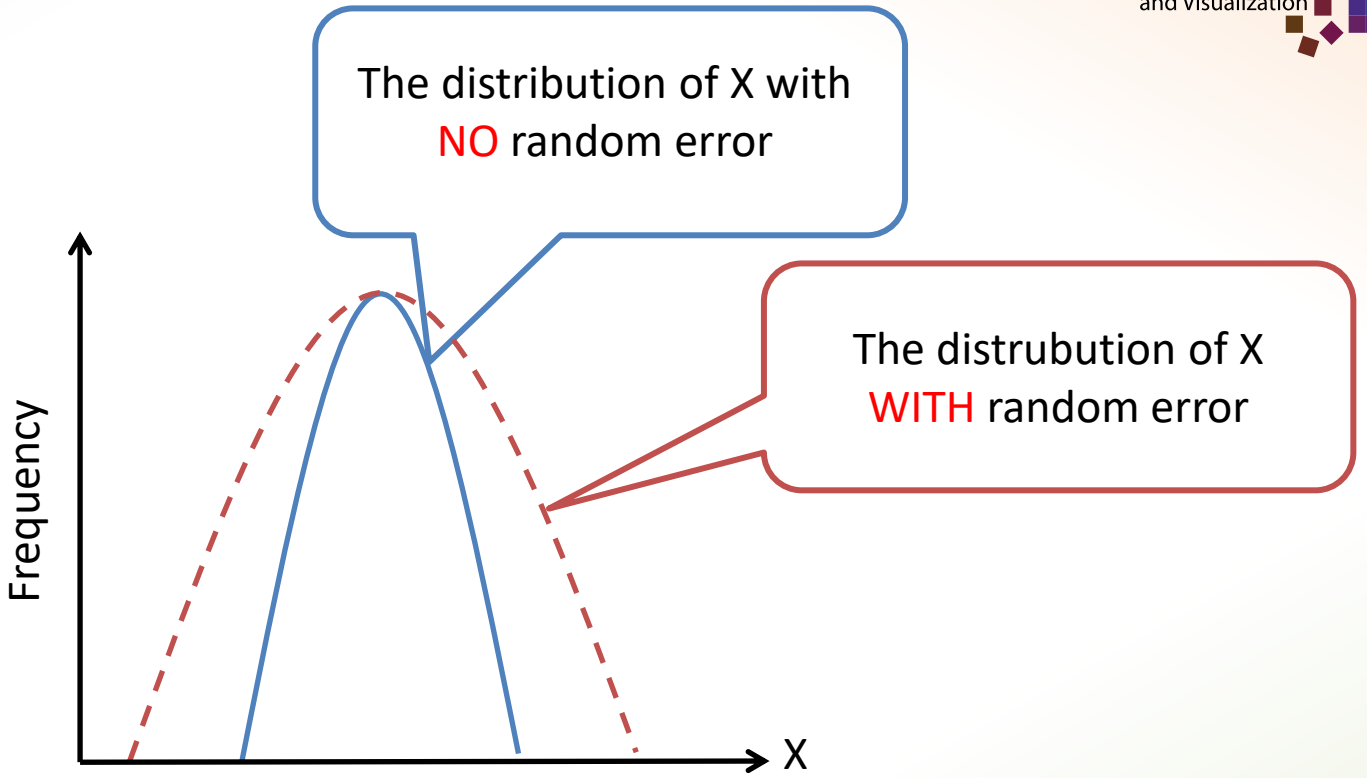
Data Cleaning: Handling Noisy Data

- **Binning -** sort data and partition into (equi-depth) bins and then smooth by bin means, bin median, bin boundaries, etc.
- **Regression -** smooth by fitting a regression function
- **Clustering –** Cluster data and remove outliers (automatically or via human inspection)

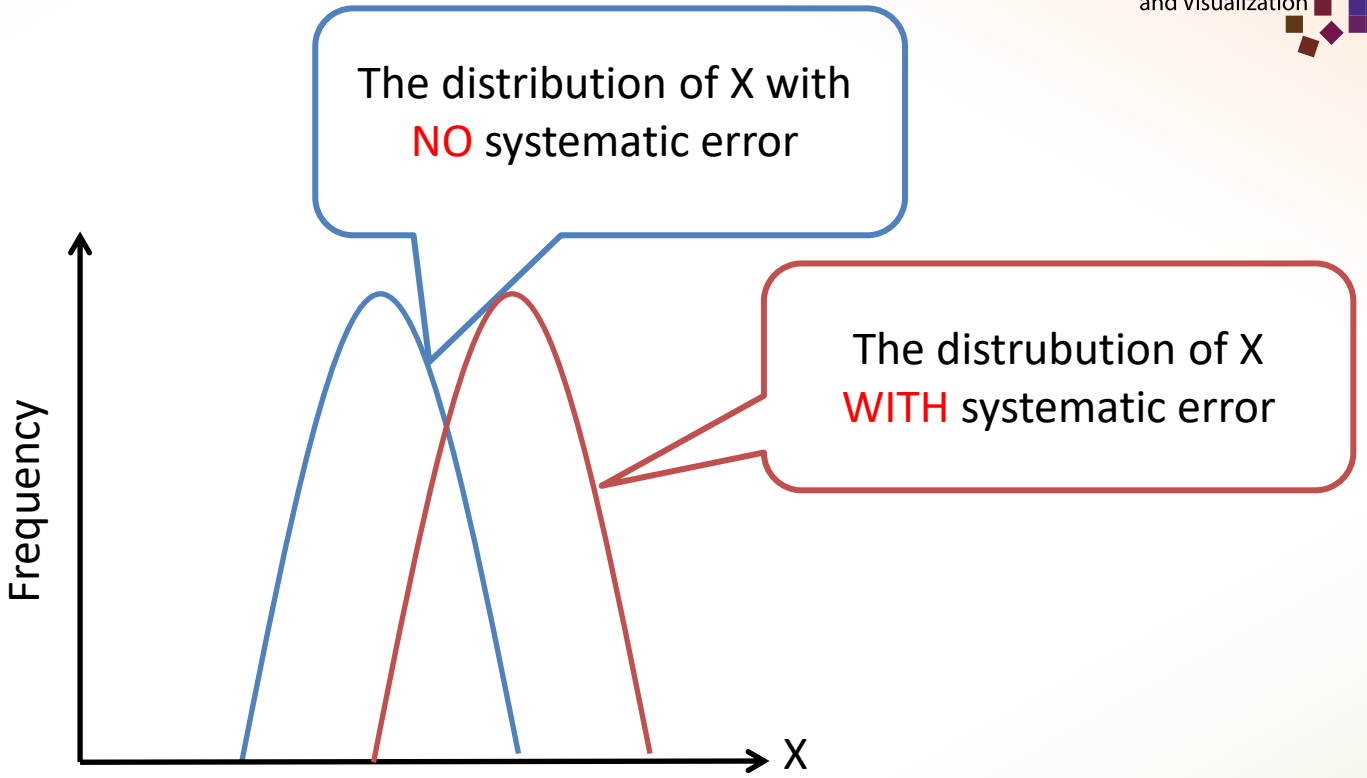
"Noise is a random error or variance in a measured variable"



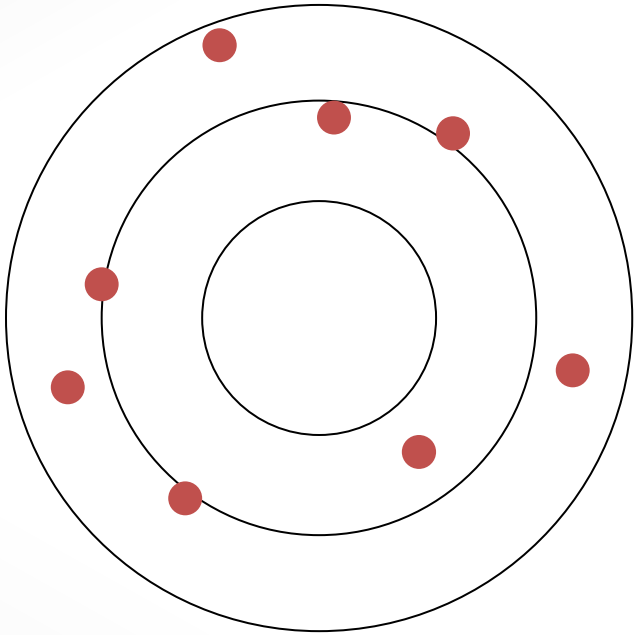
Noise is like a random error.

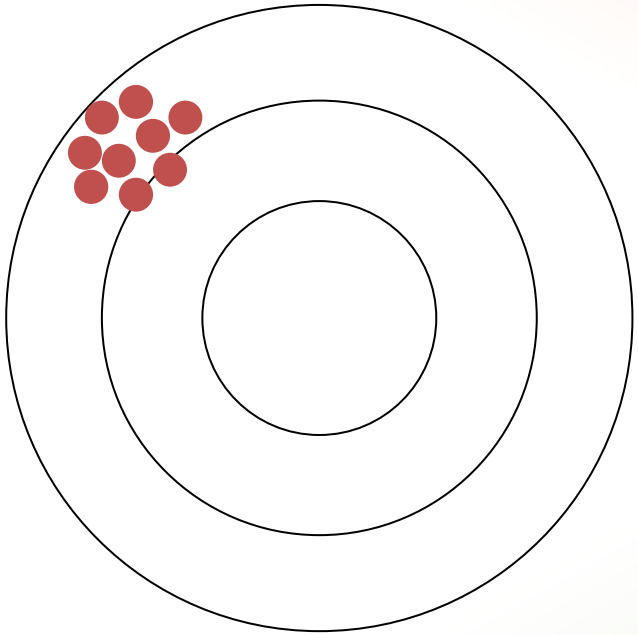


Noise is like a **random** error.



A bias is a **systematic** error.





left: random error;

right: systematic error

Data Cleaning: Handling Noisy Data

- **Binning -** sort data and partition into (equi-depth) bins and then smooth by bin means, bin median, bin boundaries, etc.
- **Regression -** smooth by fitting a regression function
- **Clustering –** Cluster data and remove outliers (automatically or via human inspection)

Data Cleaning: Handling Noisy Data

- **Binning -** sort data and partition into (equi-depth) bins and then smooth by bin means, bin median, bin boundaries, etc.
- **Regression -** smooth by fitting a regression function
- **Clustering –** Cluster data and remove outliers (automatically or via human inspection)

Handling Noisy Data: Regression

Linear regression can also be used to:

- Fill in missing values
- Classification and prediction for numeric values
- Reduce the amount of data by not storing all data values but just the model function
- Smooth data

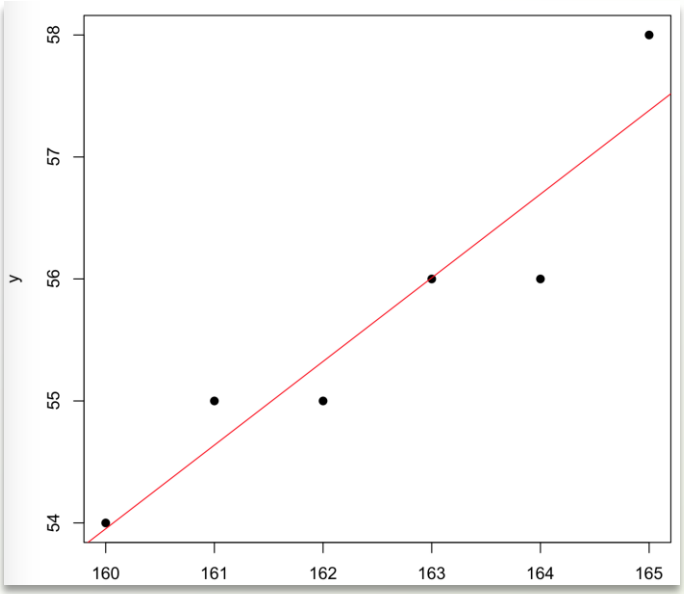
Handling Noisy Data: Regression

Smooth out noise by fitting data to a function

- Linear regression tries to discover the parameters of the straight line equation that **best fits the data** point
- **Best fits the data** = line that reduces the squared error of all data points

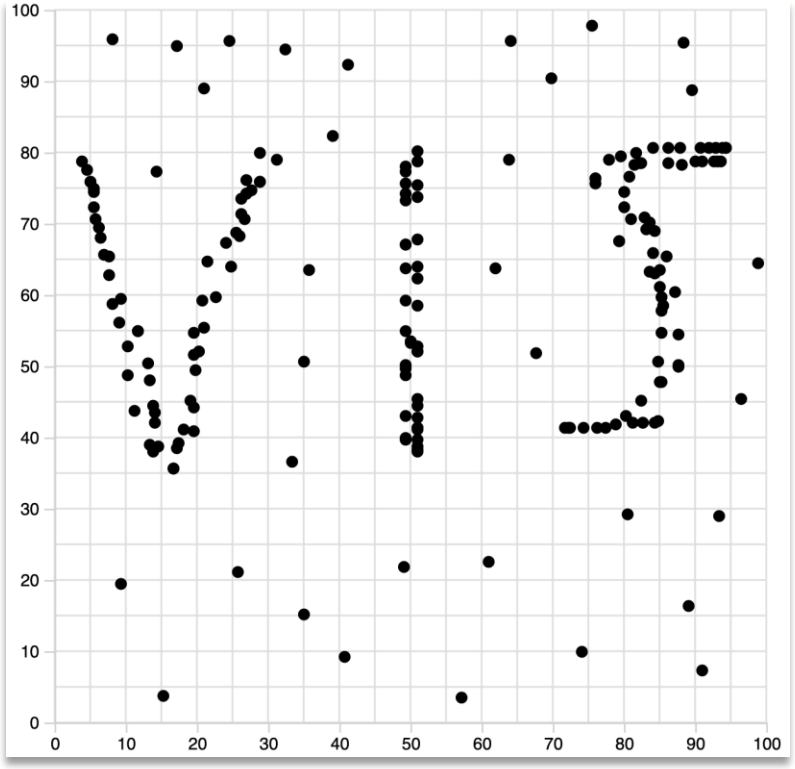
Regression (Example)

Height	x	160	161	162	163	164	165
Weight	y	54	55	55	56	56	58

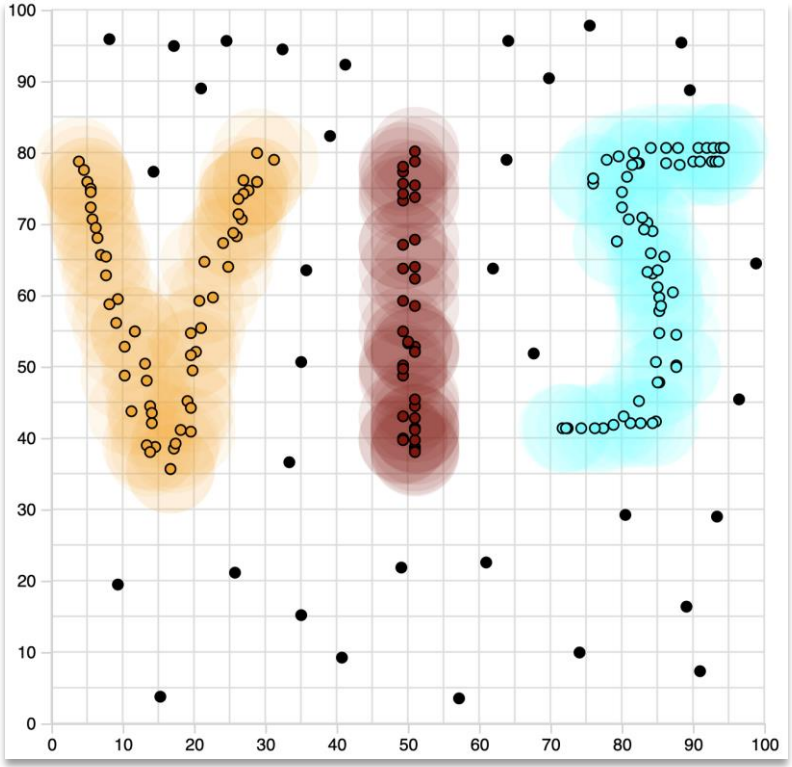


Data Cleaning: Handling Noisy Data

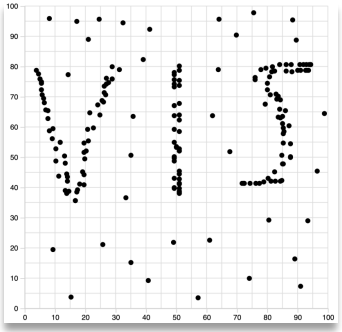
- **Binning -** sort data and partition into (equi-depth) bins and then smooth by bin means, bin median, bin boundaries, etc.
- **Regression -** smooth by fitting a regression function
- **Clustering –** Cluster data and remove outliers (automatically or via human inspection)

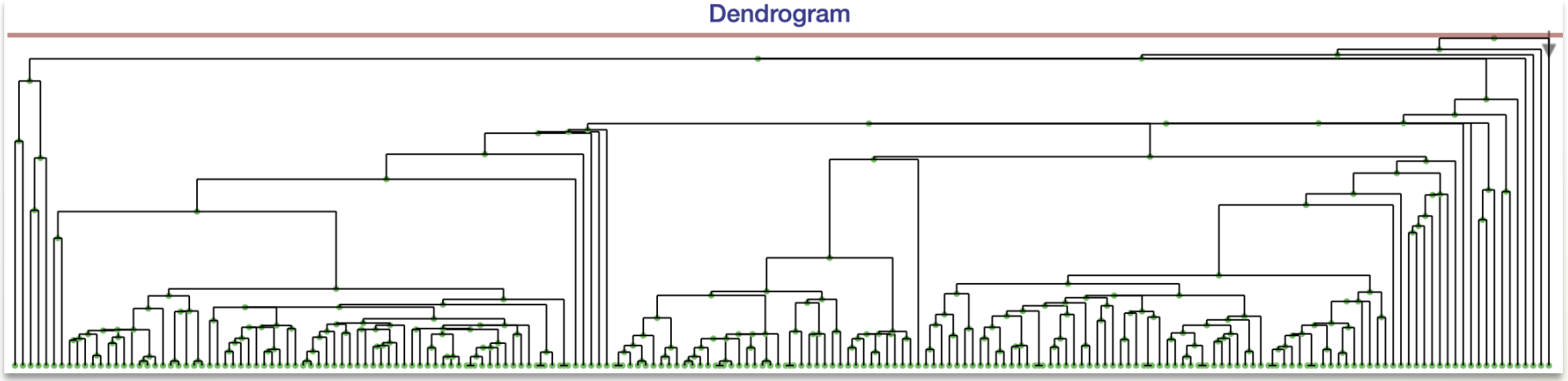


Clustering: Keep the “real” data points and filter out the noise.

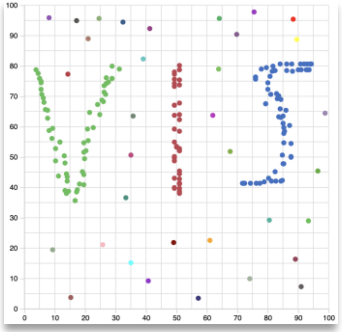


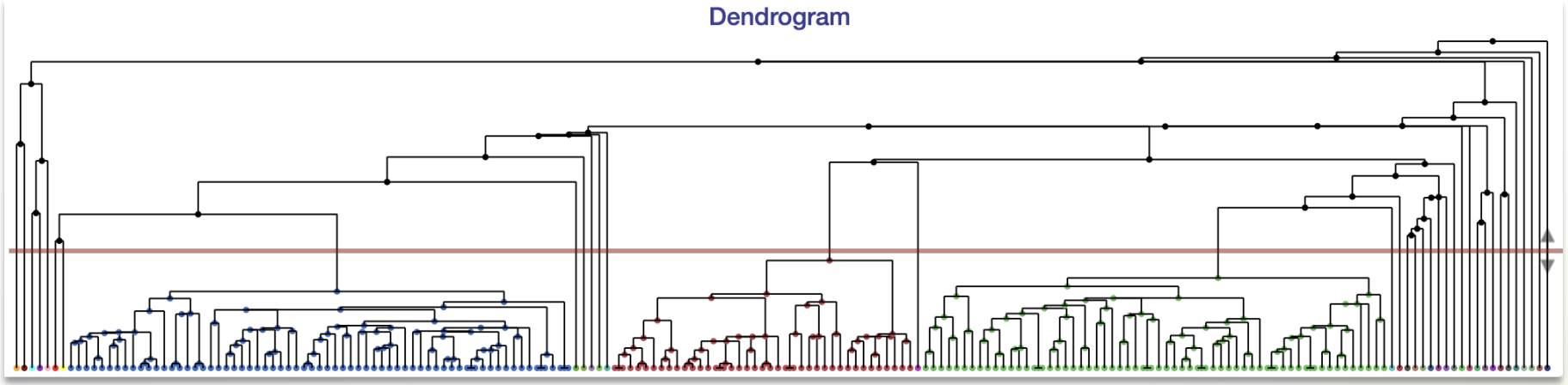
Clustering: Keep the “real” data points and filter out the noise. (DBSCAN)





Clustering: Keep the “real” data points and filter out the noise. (SINGLE-LINKAGE)





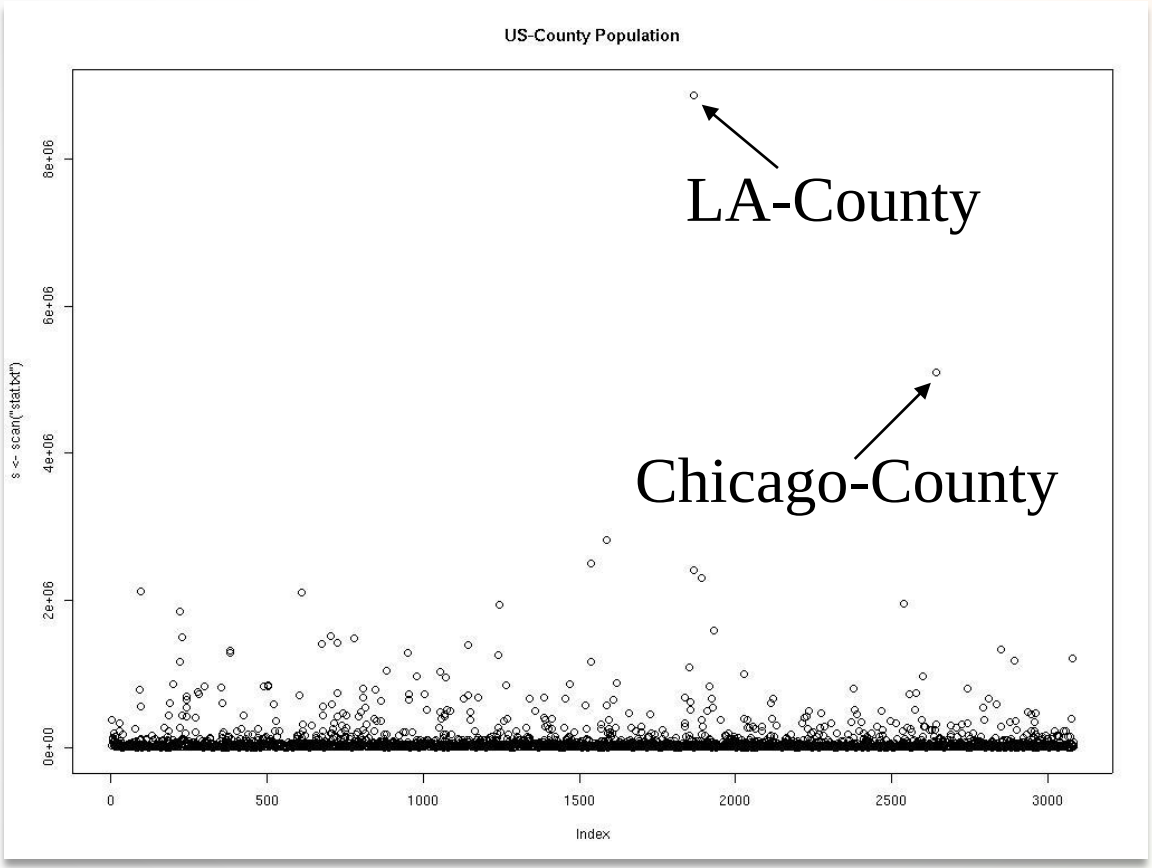
Clustering: Keep the “real” data points and filter out the noise. (SINGLE-LINKAGE)

Outline

- Types of Data
- Typical Data Classes
- Data Preprocessing
  - Data cleaning
  - **Normalization**
  - Segmentation
  - Data reduction

Normalization

- Transform features/dimensions/attributes to be on a similar scale.
  - E.g., map data to a range between 0 and 1.
- Important for some automatic calculations (e.g., clustering).
- Important for the visual mapping (e.g., enable meaningful comparisons)



Idea: represent the population density with color and display the data on a geographic map.

Normalization

- Transform features/dimensions/attributes to be on a similar scale.

$$f_{lin}(v)\frac{v\ -min}{\max - min}$$

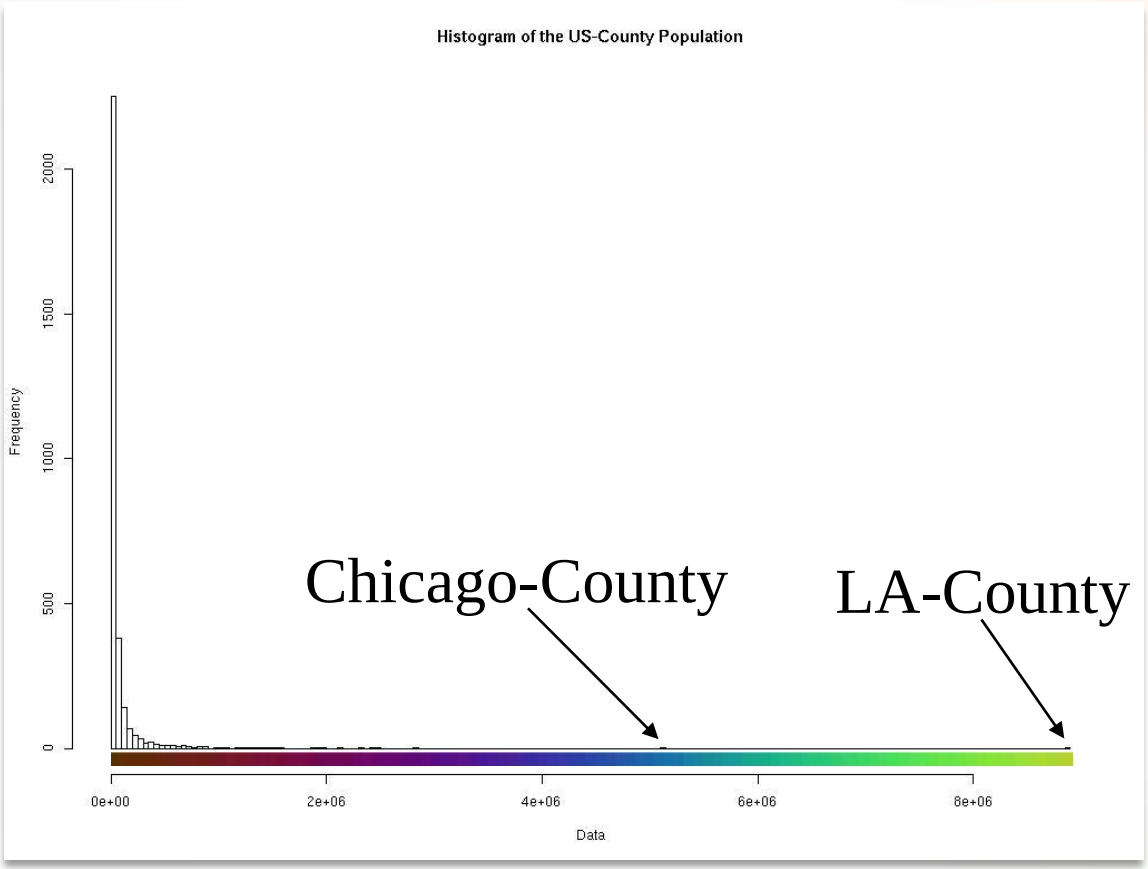
Normalization

index	1	2	3	4	5	6	7	8	9	10
data	8	7	5	6	9	2	15	11	8	6

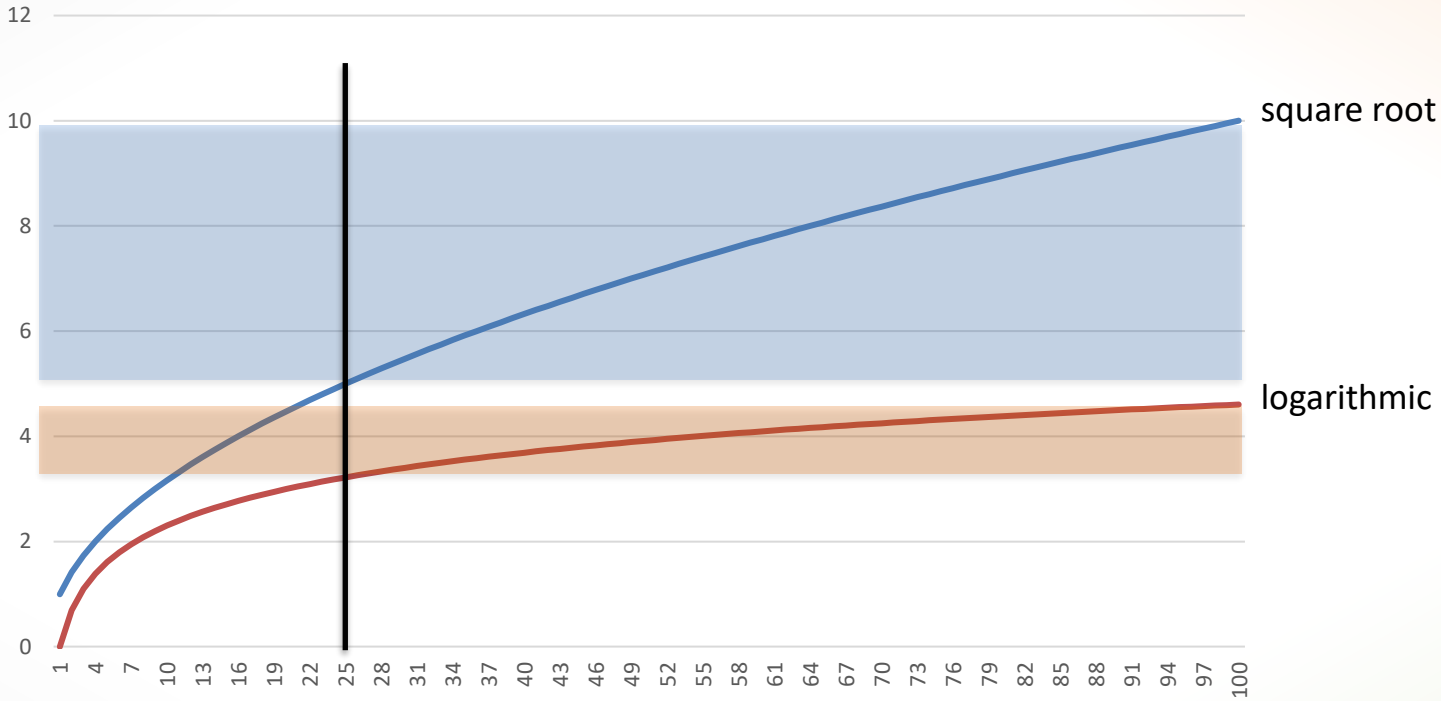
$$f_{lin}(v)\ \frac{v\ -min}{\max - min}$$

$$f_{lin}(6)\ \frac{6\ -2}{15-2} = 0.31$$

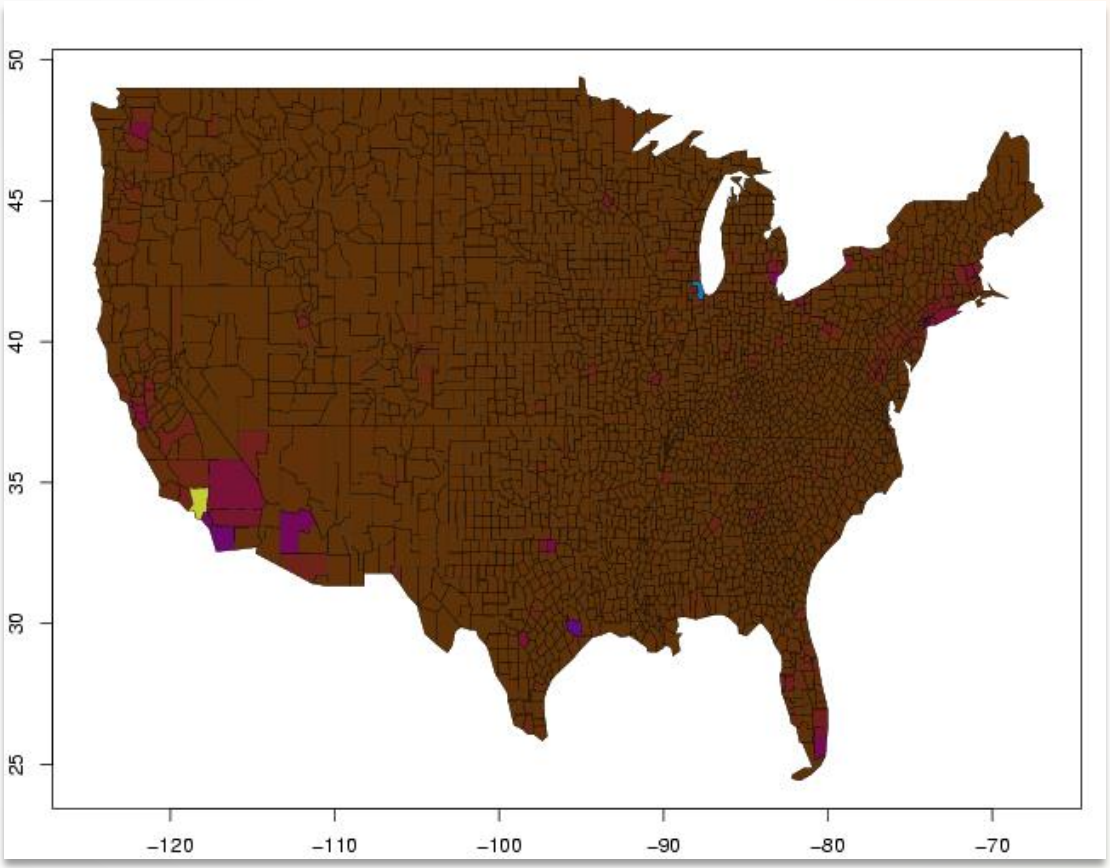
index	1	2	3	4	5	6	7	8	9	10
data	8	7	5	6	9	2	15	11	8	6
L-Norm	.46	.38	.23	.31	.54	0	1	.69	.46	.31



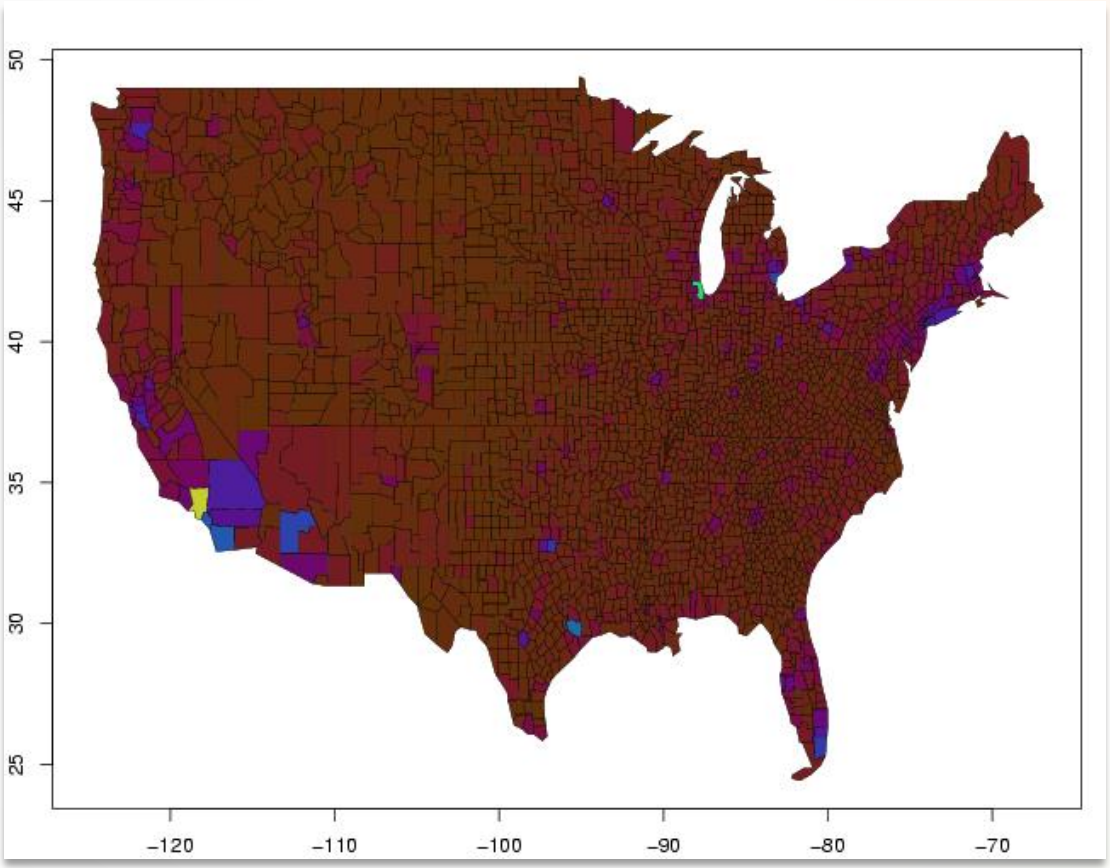
Idea: represent the population density with color and display the data on a geographic map.



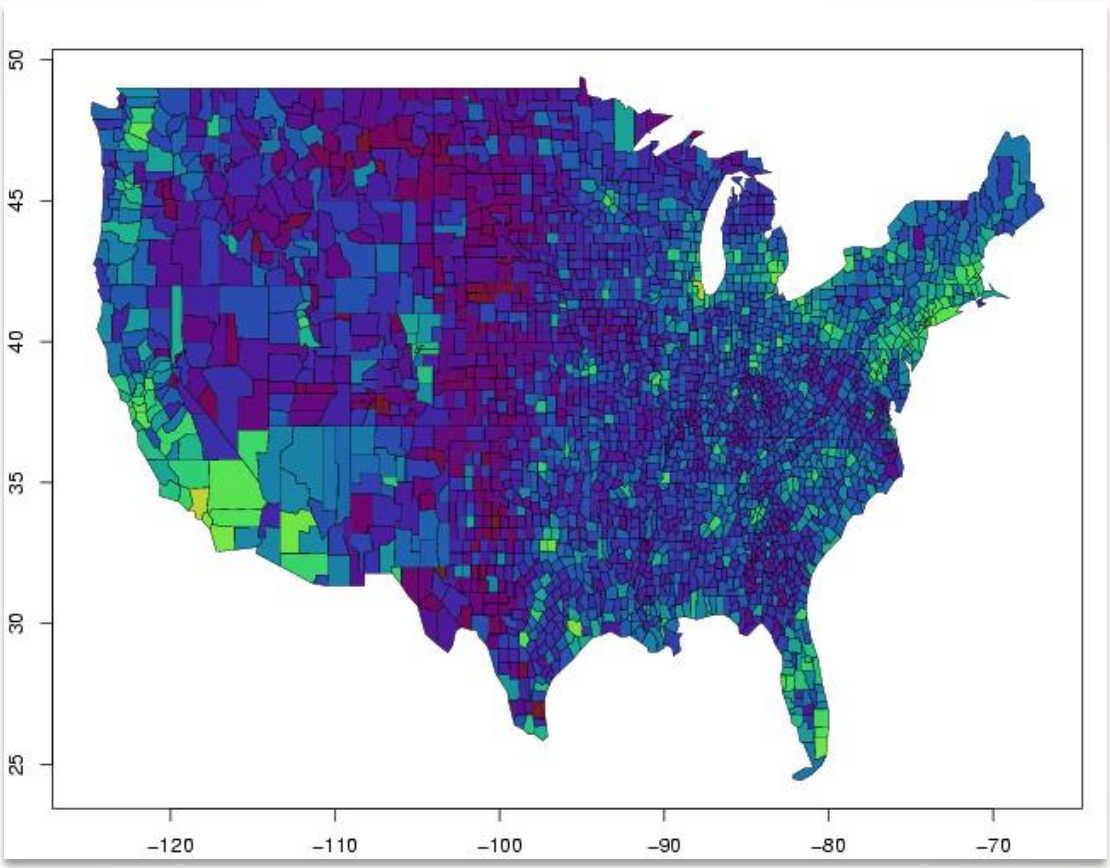
Logarithmic vs square root normalization.



Linear normalization of the population density.



Square root normalization of the population density.



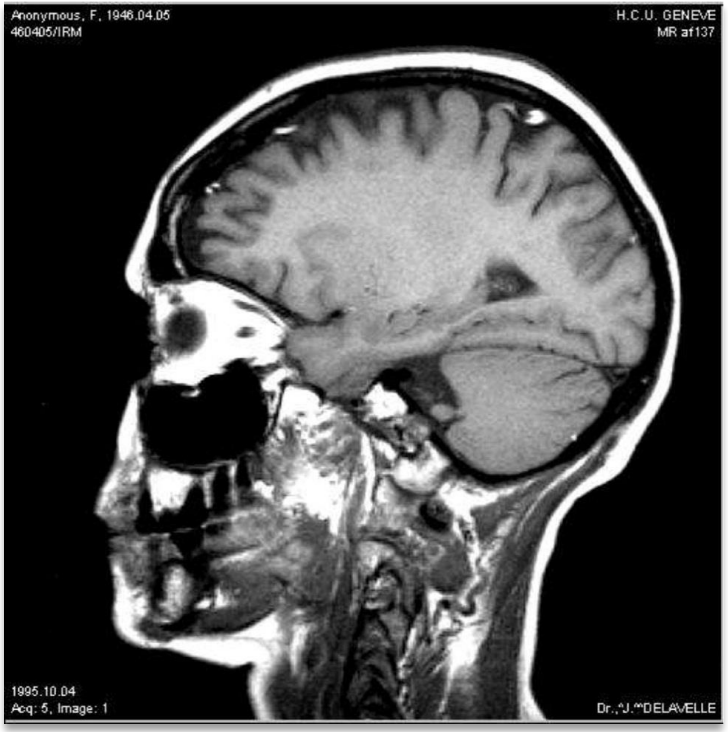
Logarithmic normalization of the population density.

Outline

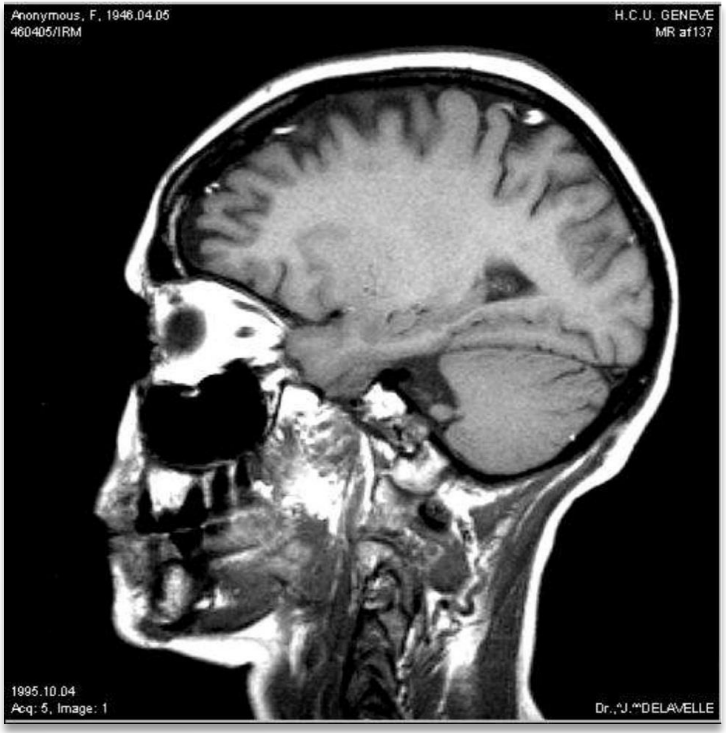
- Types of Data
- Typical Data Classes
- Data Preprocessing
  - Data cleaning
  - Normalization
  - **Segmentation**
  - Data reduction

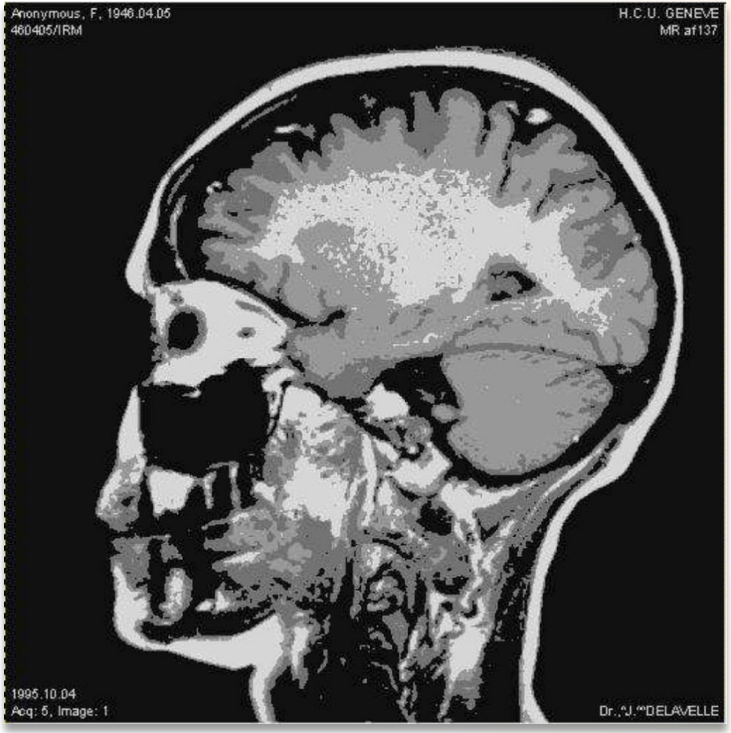
Segmentation: Manual

- Simplify a visualization into something more meaningful and easier to analyze.
- Usually, by applying thresholds for bins interactively.
  - Grouping of similar data points
- Especially useful if there are predefined categories.



MRI scan: it is interesting to distinguish between bones, muscles, fat etc.

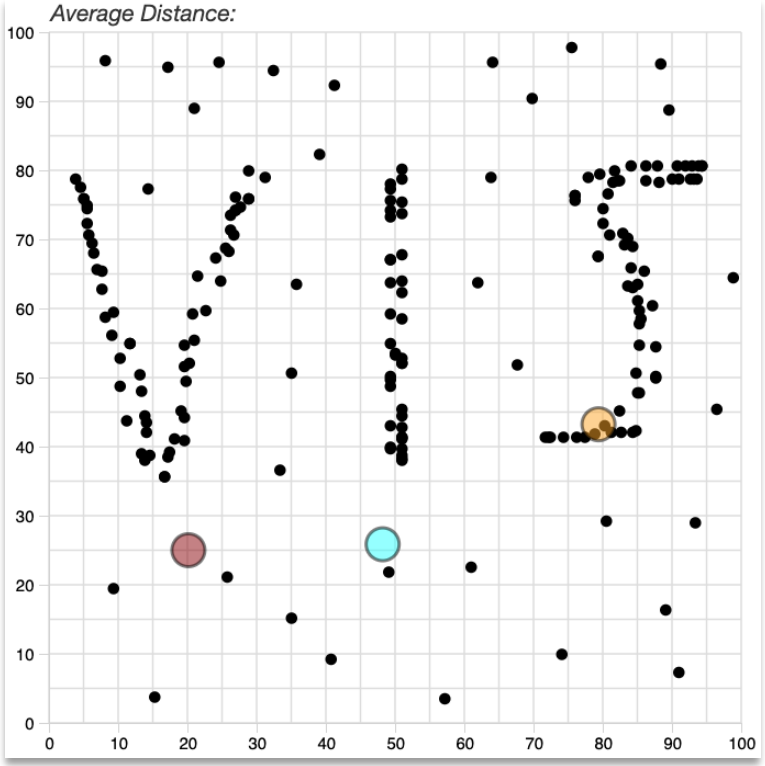




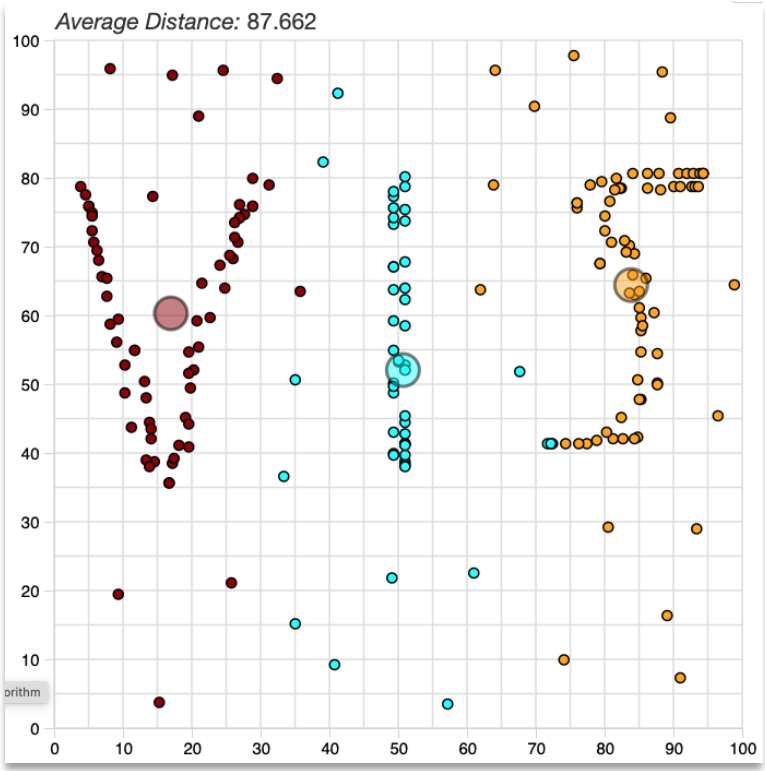
Simplified MRI scan: no continuous color values but categories.

Segmentation: Automatic

- Automatically segment data points.
  - Grouping of similar data points.
- Clustering is the process of finding similarities in the data (distances, densities, etc.).
- Therefore, clustering techniques can be used to automatically segment data.



Automatic segmentation of data points using k-means clustering.



Automatic segmentation of data points using k-means clustering.

Outline

- Types of Data
- Typical Data Classes
- Data Preprocessing
  - Data cleaning
  - Normalization
  - Segmentation
  - **Data reduction**

Data Reduction

- **Reduction of number of data points.**
- Reduction of number of dimensions.

Data Reduction: Sampling

- Large data sets may take data mining algorithms excessive amounts of time and/or memory to process.
- Data may be expensive to collect, so a representative subset is needed.

Goal:

- The subset should sufficiently represent the whole data set

Data Reduction: Sampling

Types of sampling:

- Non-probabilistic sampling
  - Sample selected on some non-random basis (volunteers, female,...)
- Probabilistic sampling
  - Sample selected on the basis of random selection so that every element of the data set has an equal chance of being selected.

Data Reduction

- Reduction of number of data points.
- **Reduction of number of dimensions.**
  - **Feature Selection**
  - **Feature Extraction**

Dimension Reduction

- Facilitate the design of models (machine learning).
  - Easier to train.
  - Easier to interpret.
  - Easier to generalize -> less likely overfitted.
  - …
- Reduces the complexity of the data and, therefore, facilitates the mapping to visual features (data visualization).
  - Improves the scalability.
  - Easier to detect patterns.

Dimension Reduction: Feature Selection

- Manually select features of interest (background knowledge).

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21	6	160	110	3,9	2,62	16,46	0	1	4	4
Mazda RX4 Wag	21	6	160	110	3,9	2,875	17,02	0	1	4	4
Datsun 710	22,8	4	108	93	3,85	2,32	18,61	1	1	4	1
Hornet 4 Drive	21,4	6	258	110	3,08	3,215	19,44	1	0	3	1
Hornet Sportabout	18,7	8	360	175	3,15	3,44	17,02	0	0	3	2
Valiant	18,1	6	225	105	2,76	3,46	20,22	1	0	3	1
Duster 360	14,3	8	360	245	3,21	3,57	15,84	0	0	3	4

The data was extracted from the 1974 *Motor Trend* US magazine

Dimension Reduction: Feature Selection

- Automatic feature selection (e.g., information gain).

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21	6	160	110	3,9	2,62	16,46	0	1	4	4
Mazda RX4 Wag	21	6	160	110	3,9	2,875	17,02	0	1	4	4
Datsun 710	22,8	4	108	93	3,85	2,32	18,61	1	1	4	1
Hornet 4 Drive	21,4	6	258	110	3,08	3,215	19,44	1	0	3	1
Hornet Sportabout	18,7	8	360	175	3,15	3,44	17,02	0	0	3	2
Valiant	18,1	6	225	105	2,76	3,46	20,22	1	0	3	1
Duster 360	14,3	8	360	245	3,21	3,57	15,84	0	0	3	4

The data was extracted from the 1974 *Motor Trend* US magazine

Dimension Reduction: Feature Extraction

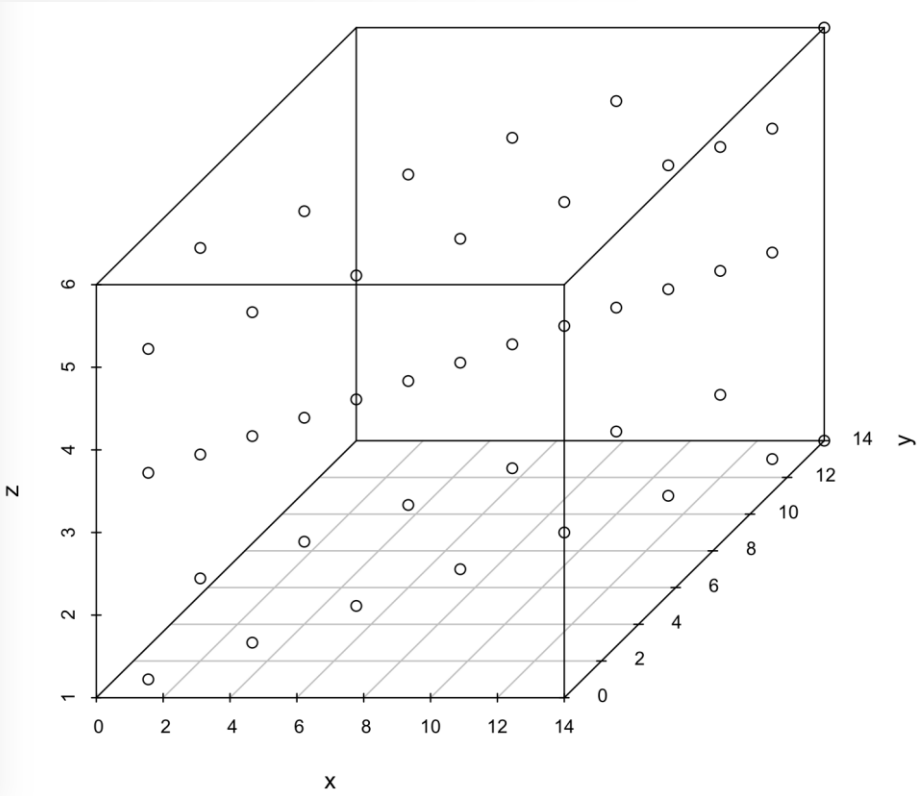
- Based on the given features, new features are derived.
  - The number of features depends on the technique used
    - PCA: number of derived features == number of given features.
    - MDS: number of derived features is an input parameter.
    - UMAP: number of derived features is an input parameter.
    - T-SNE: number of derived features is an input parameter.
    - …
- Sometimes these new features are not interpretable
  - E.g., MDS, UMAP, T-SNE

Dimension Reduction: Feature Extraction

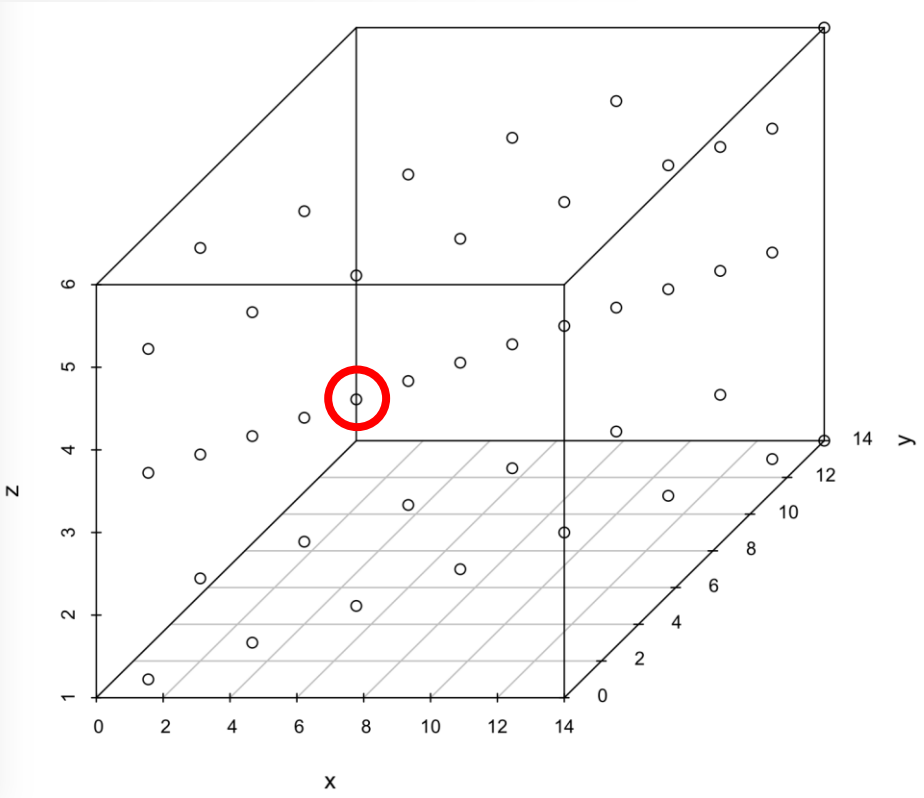
- It depends on the method used, whether the new features are interpretable
  - The new derived features ARE interpretable
    - PCA: linear combination of the given features
  - The new derived features ARE NOT interpretable
    - MDS: used to somehow show distances between data points.
    - UMAP: used to provide space for projecting a force-directed graph.
    - T-SNE: used to group data points based on probabilities.

Feature Extraction: PCA

- What is a **P**rincipal **C**omponent **A**nalysis doing:
  - Explore statistical **correlations** among elements of the original patterns.
  - Find data representation that retains the maximum **nonredundant** and **uncorrelated** information.



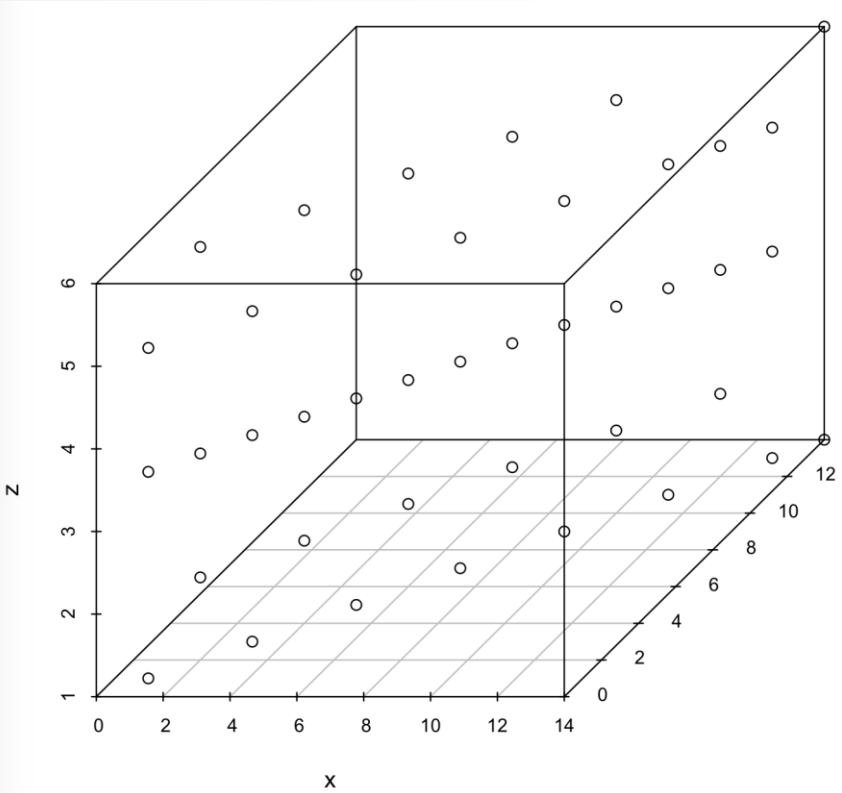
PCA: Example with three dimensions.

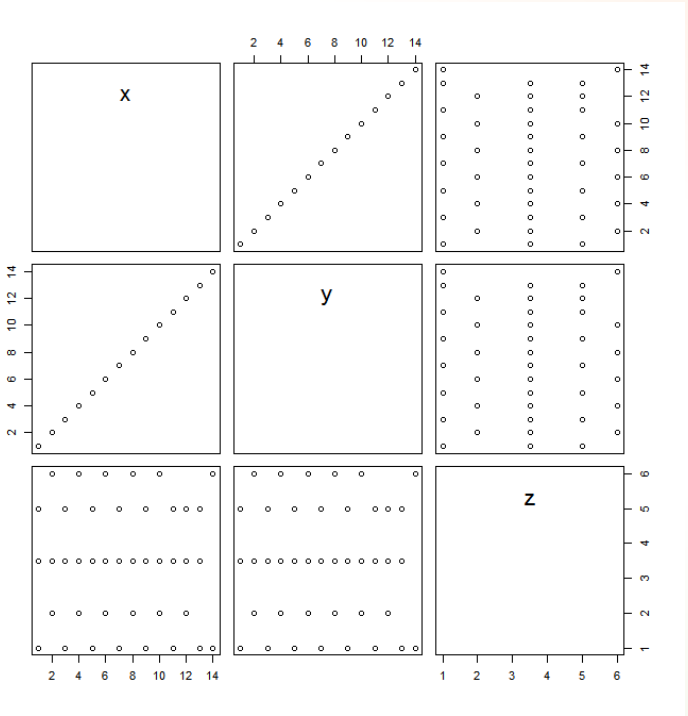


Difficult to read the exact coordinates

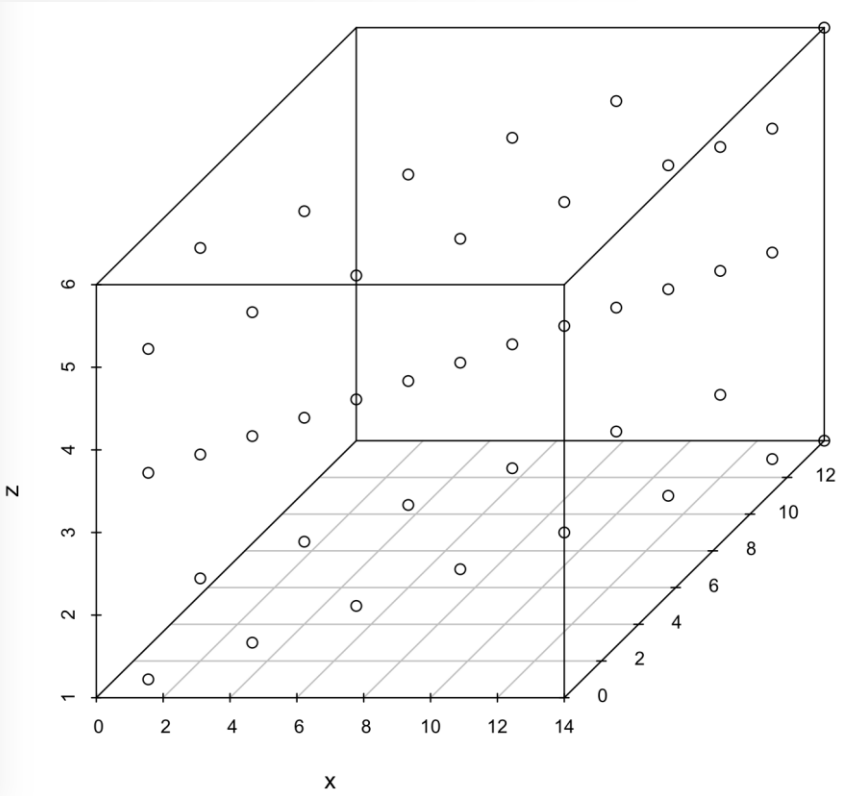
PCA: Example with three dimensions.

Scatterplot Matrix

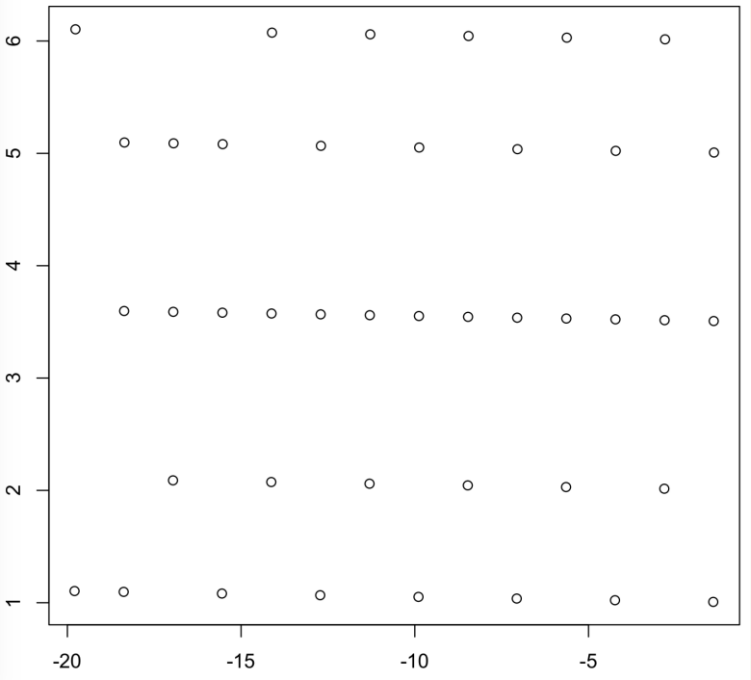




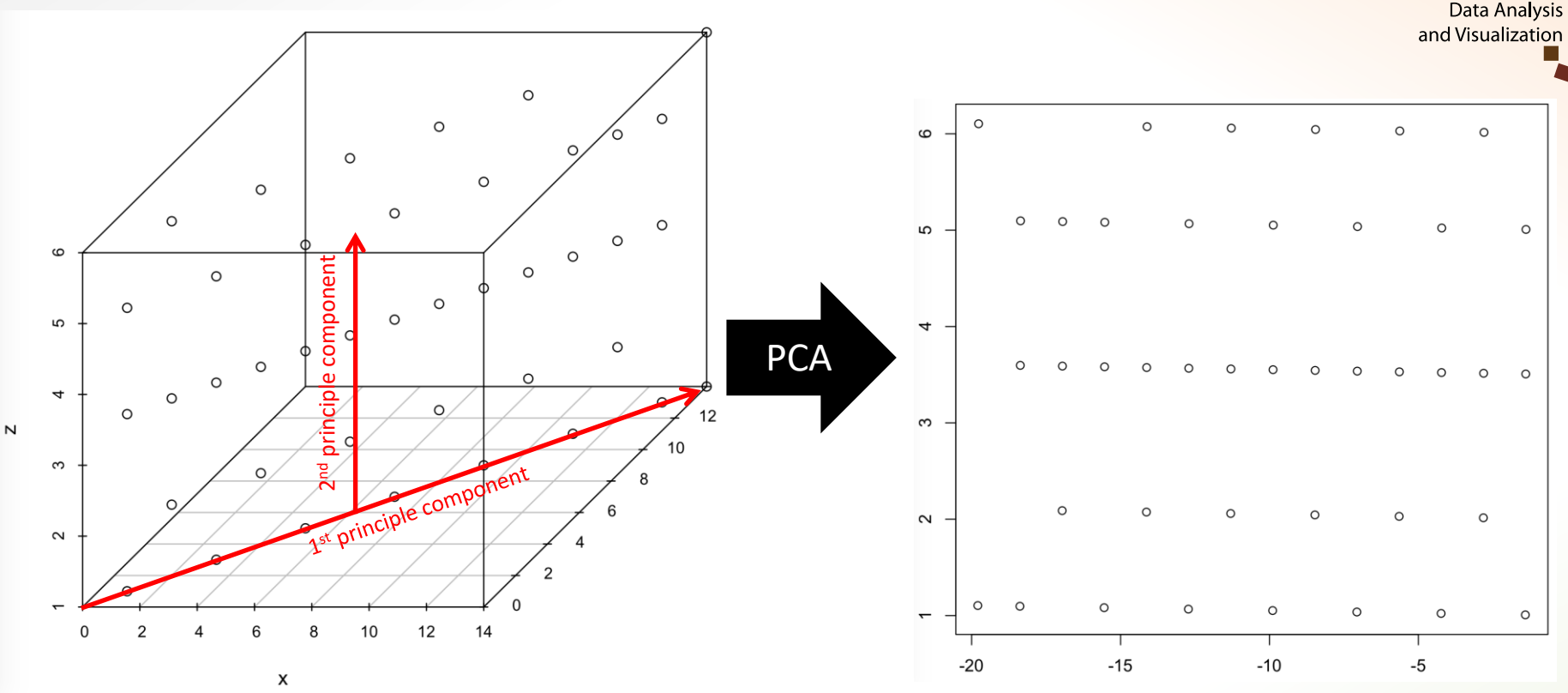
PCA: Example with three dimensions.



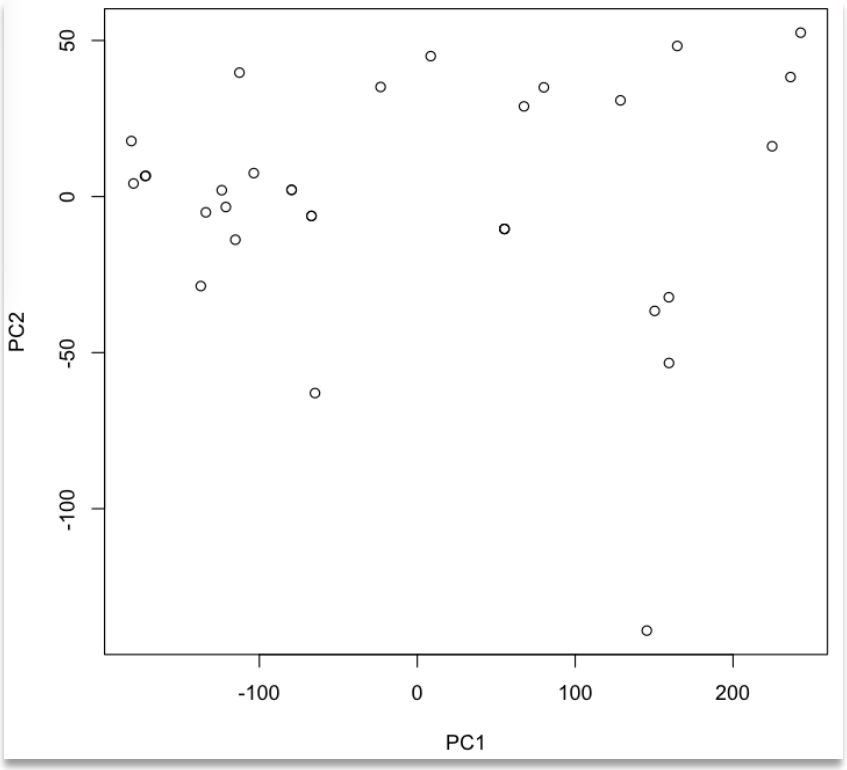
PCA



PCA: Example with three dimensions.

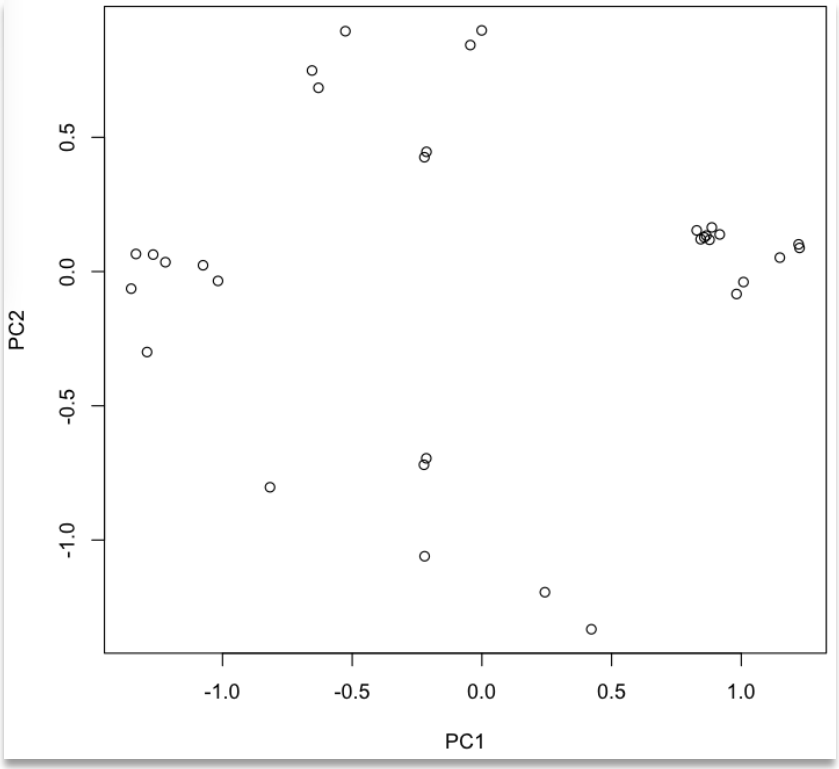


PCA: Example with three dimensions.

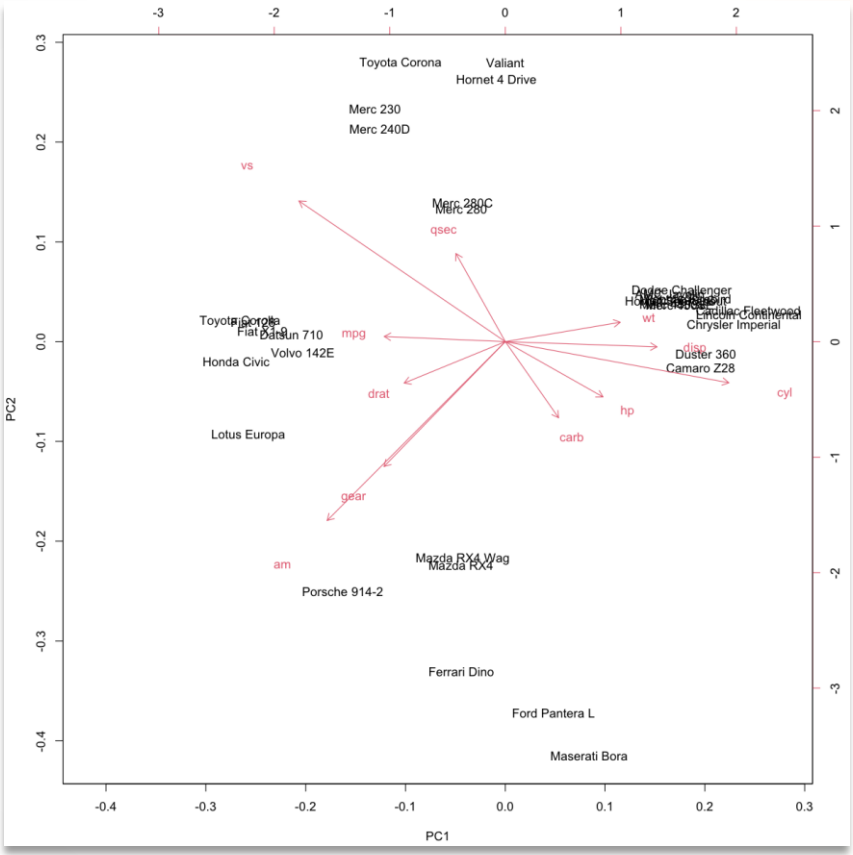


PCA projection of the mtcars dataset (available in R).

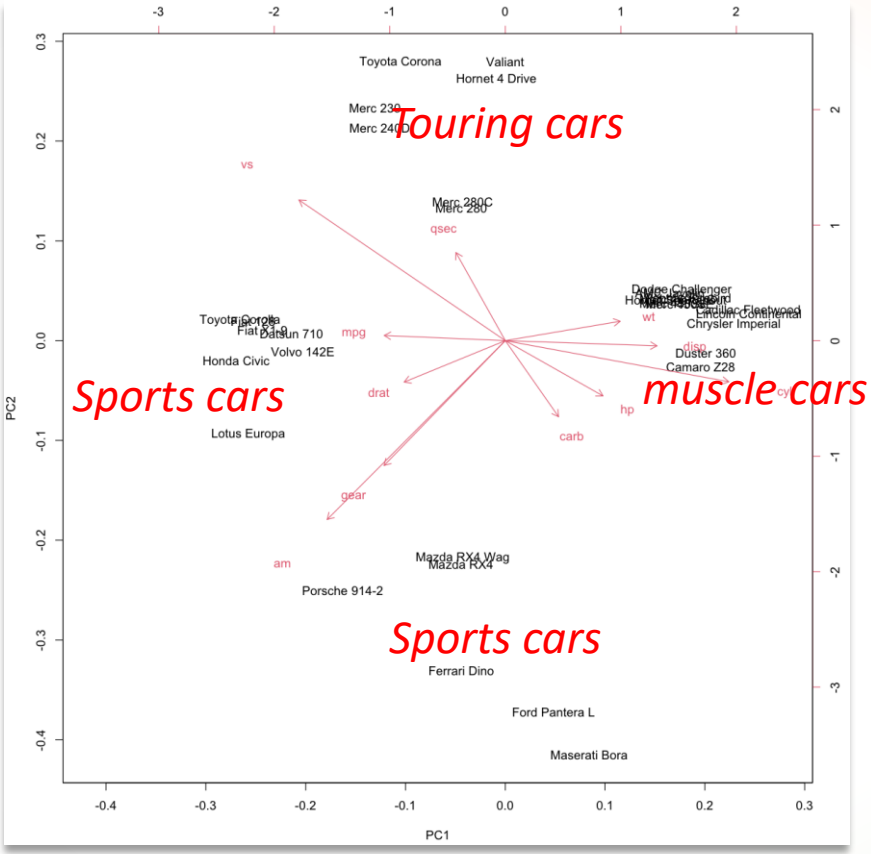
The data was extracted from the 1974 *Motor Trend* US magazine



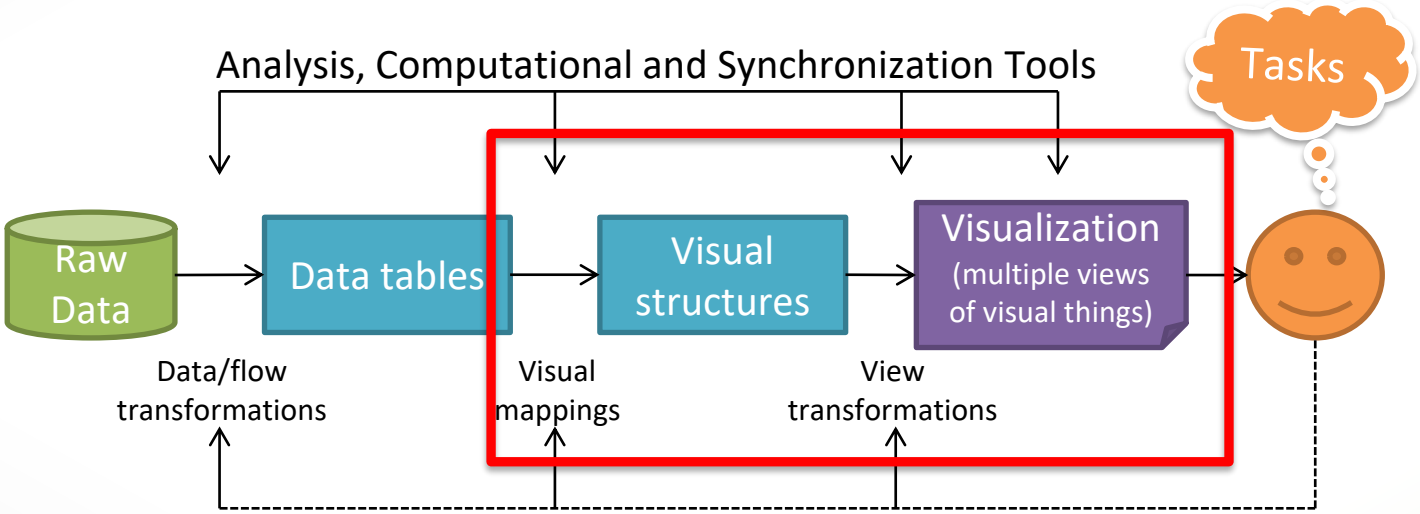
PCA projection of the mtcars dataset (available in R) – after normalization.



PCA: visualize the contribution of the original dimensions.

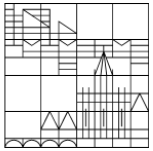


PCA: visualize the contribution of the original dimensions.



Information Visualization Reference Model

Universität Konstanz



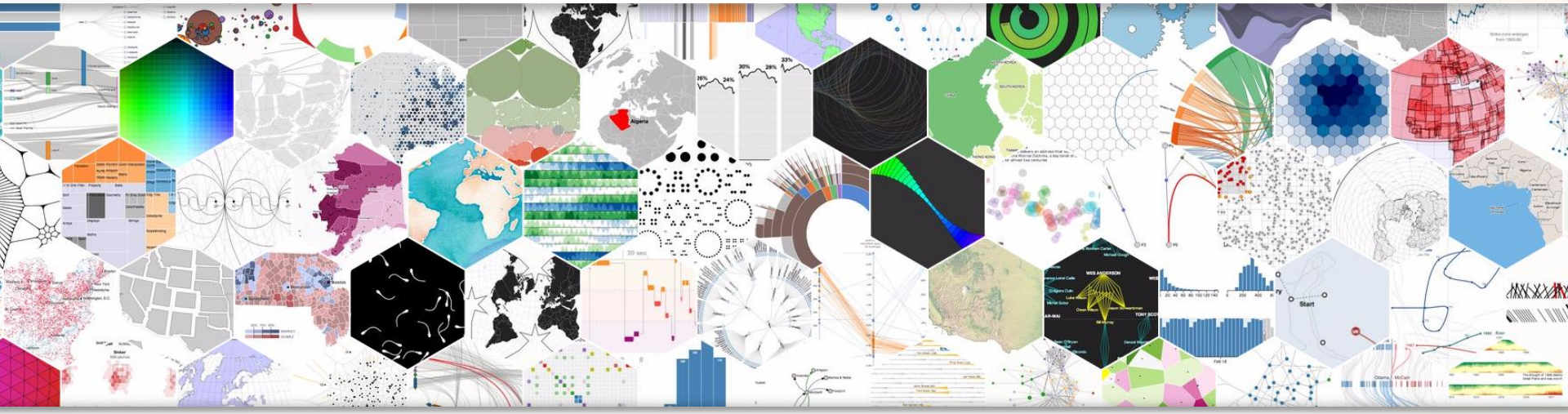
Data Analysis and Visualization

Data Visualization

Visualization Foundation

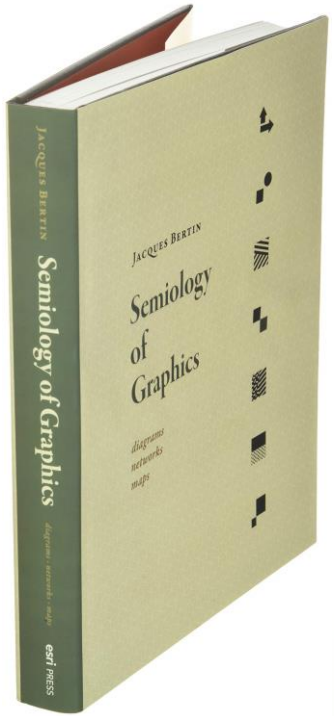
Guiding Questions

- How can we map data points to visual structures?
- Are there more suitable mappings?
- Can we evaluate a visualization?

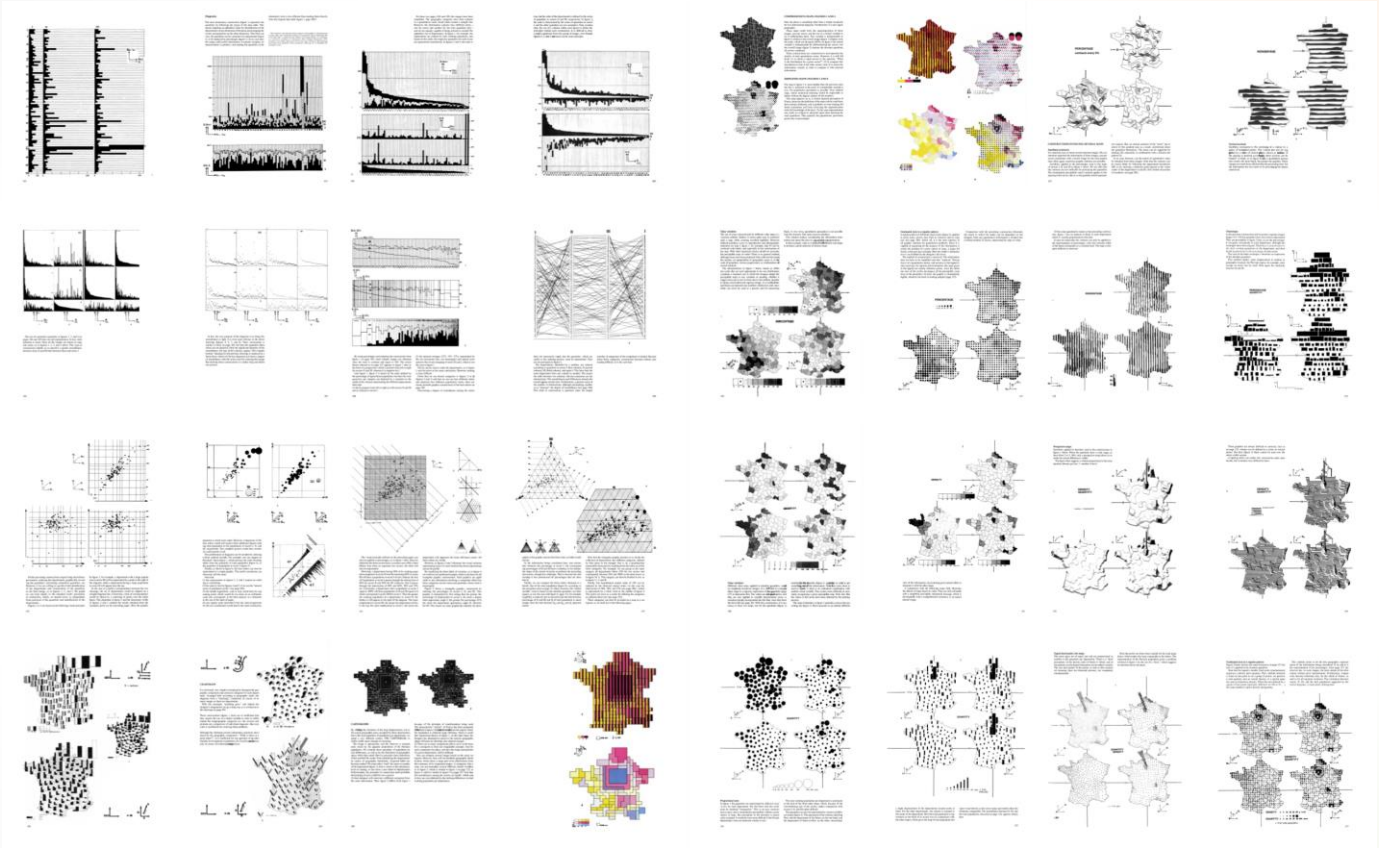


Decomposing Visualizations

Visual Language is a Sign System



- Images are perceived as a set of signs.
- Sender encodes information in signs.
- Receiver decodes information from signs.



Bertin became a pioneer in information visualization.

Visual Language is a Sign System

Strict separation of:

- **Content** (i.e., the information to encode)
- **Container** (i.e., the properties of the graphical system)

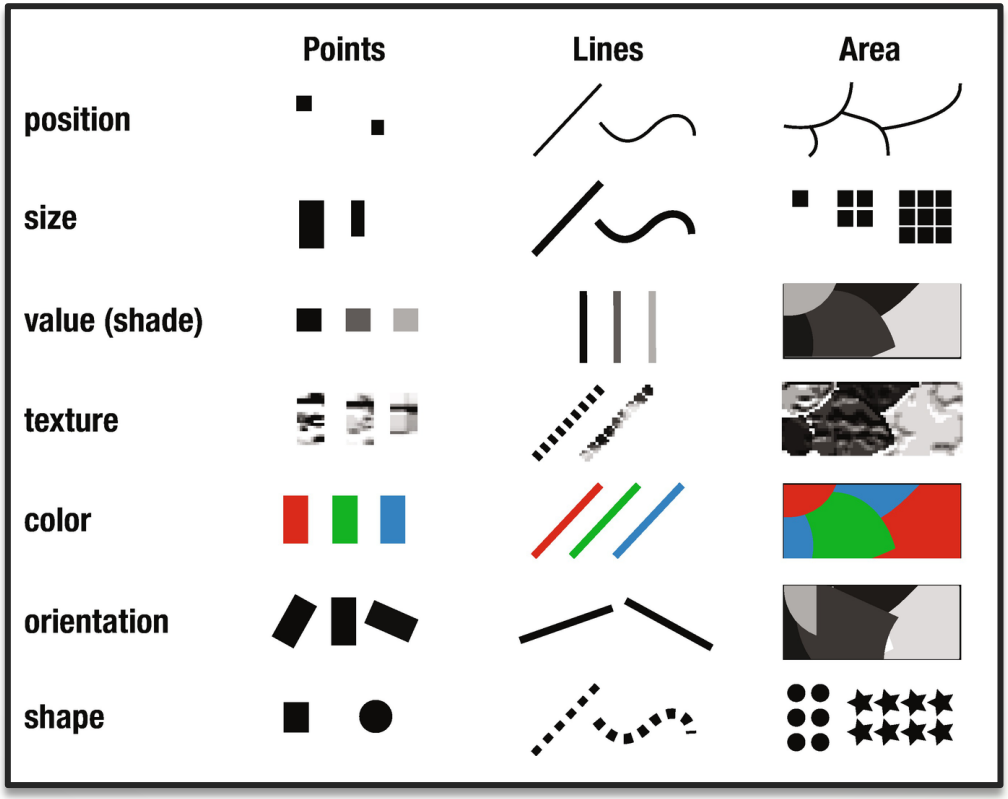
Graphics are a combination of:

- **Plane properties** (i.e., position of marks)
- **Retinal variables** (i.e., visual variables above the plane)

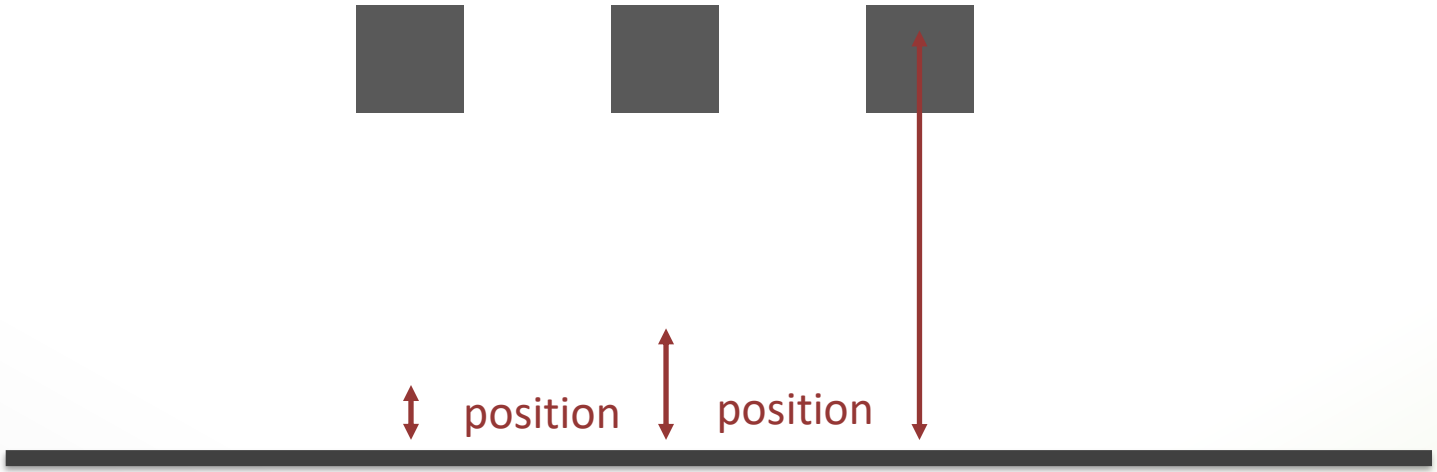
Semiology of Graphics

Bertin's graphical vocabulary:

- Marks:
  - Points, lines, and areas
- Positional:
  - Two planar dimensions
- Retinal:
  - Size, value (brightness), texture, color, orientation, shape

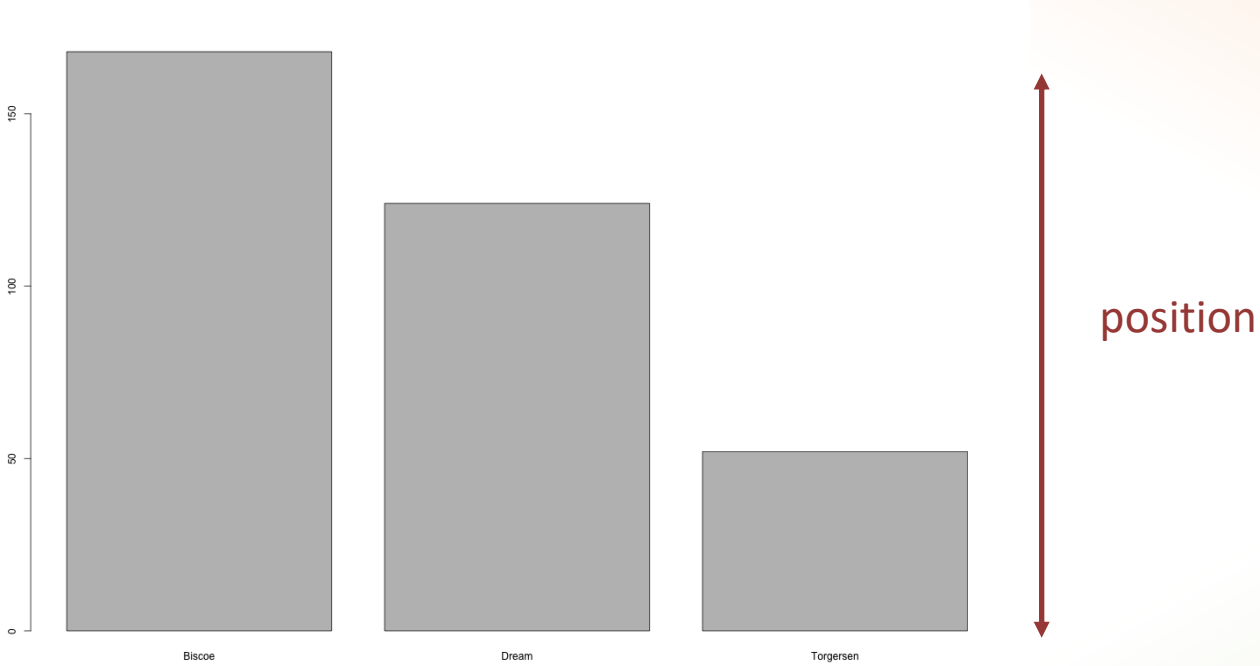


Combination of marks with positional and retinal variables.

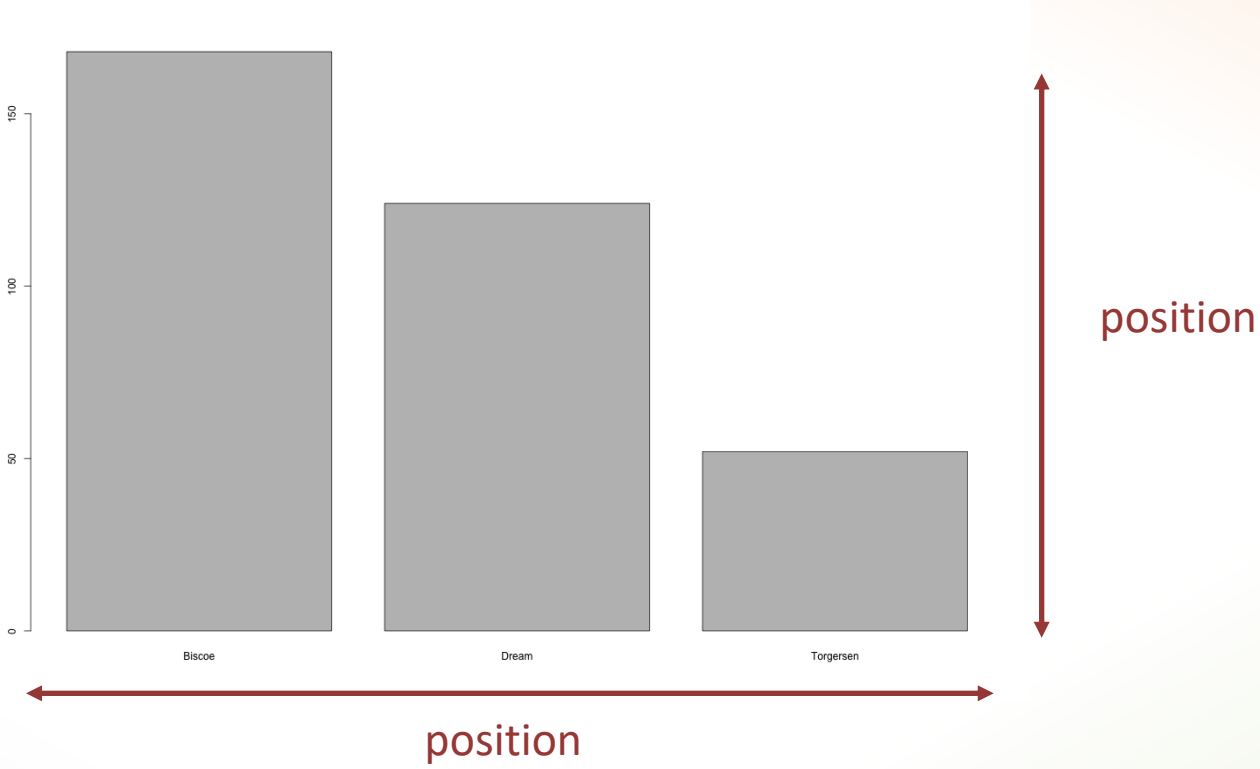


An identical baseline for comparison tasks.

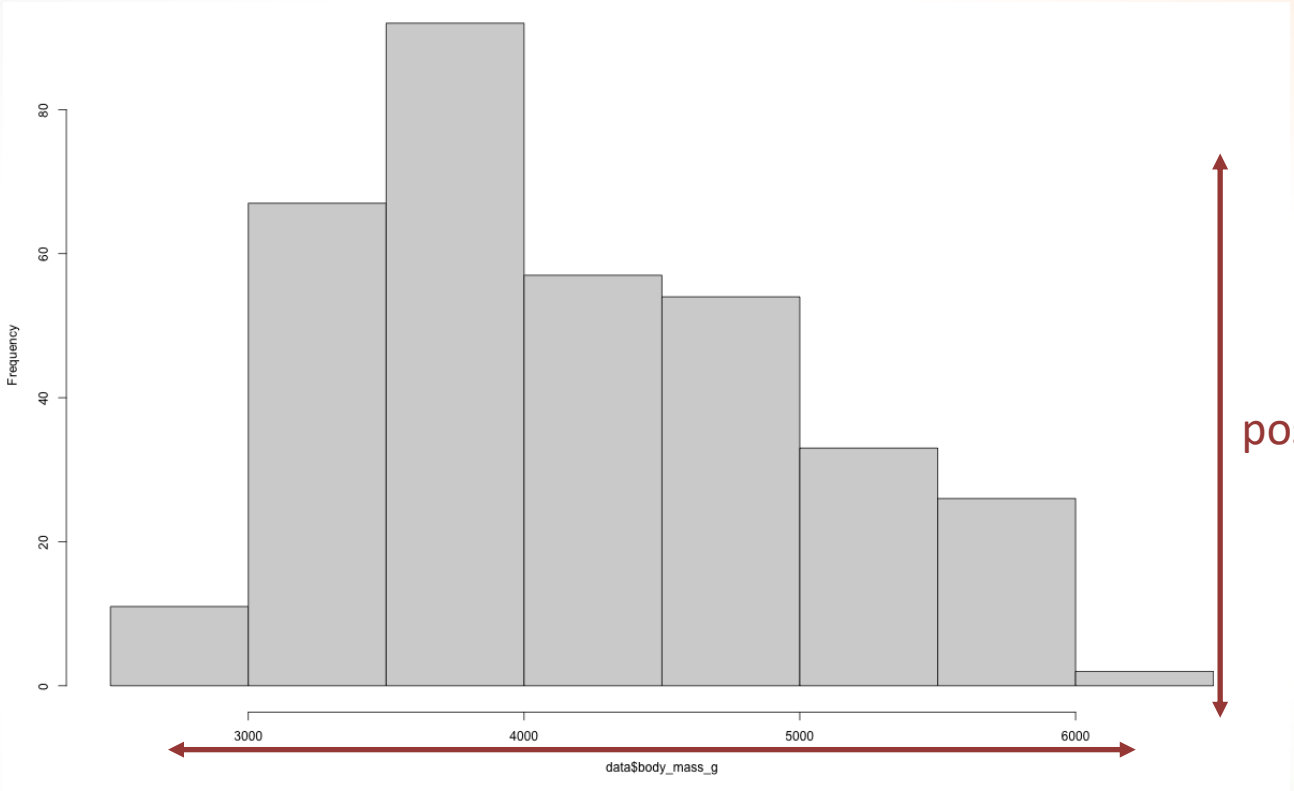
Using **POSITION** to convey information.



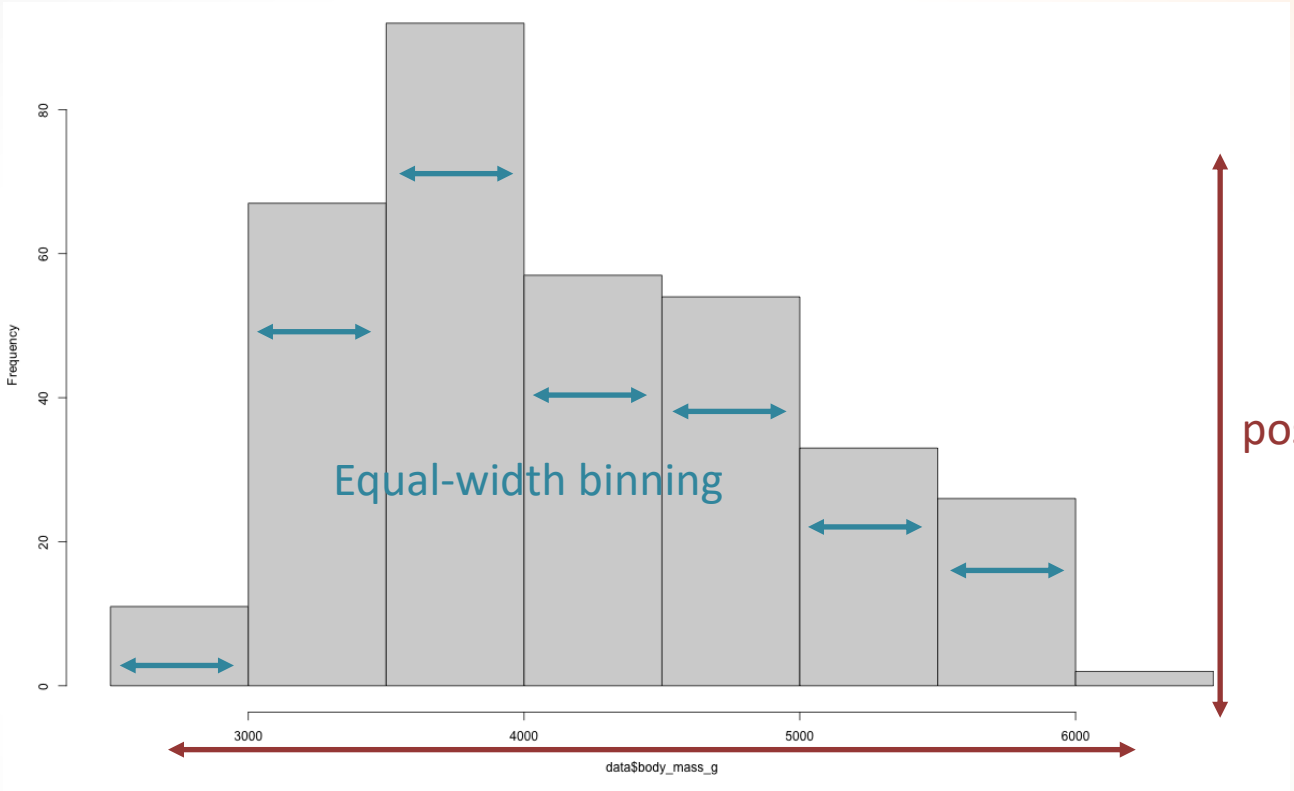
Bar chart representing Penguins' living spaces (islands).



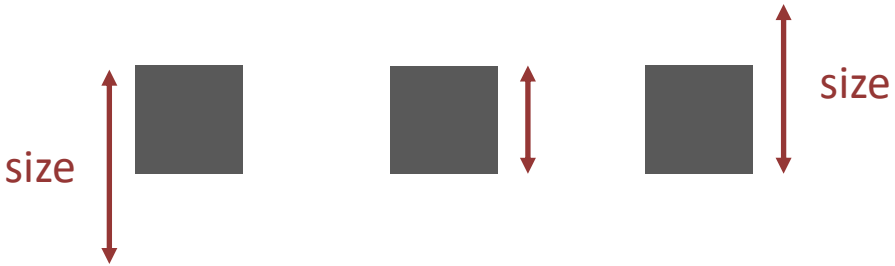
Bar chart representing Penguins' living spaces (islands).



Histogram of penguins' body mass index.

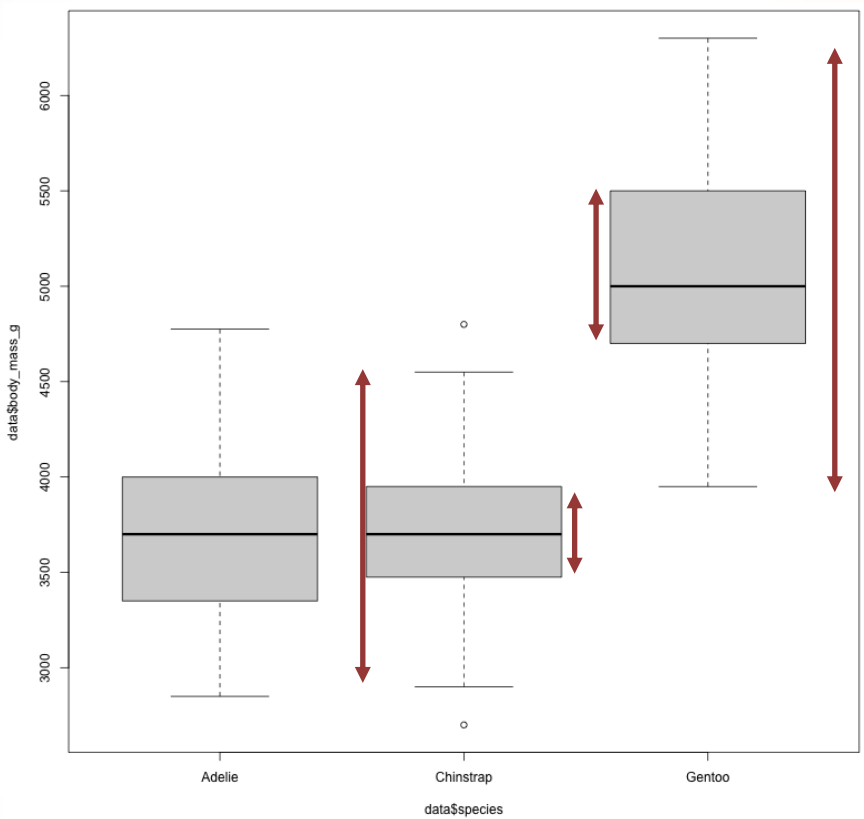


Histogram of penguins' body mass index.

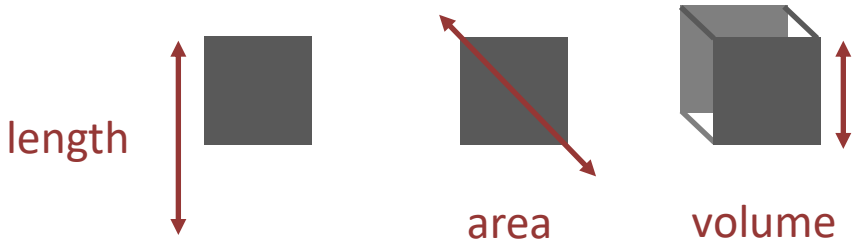


A different baseline for comparison tasks.

Using **SIZE** to convey information.

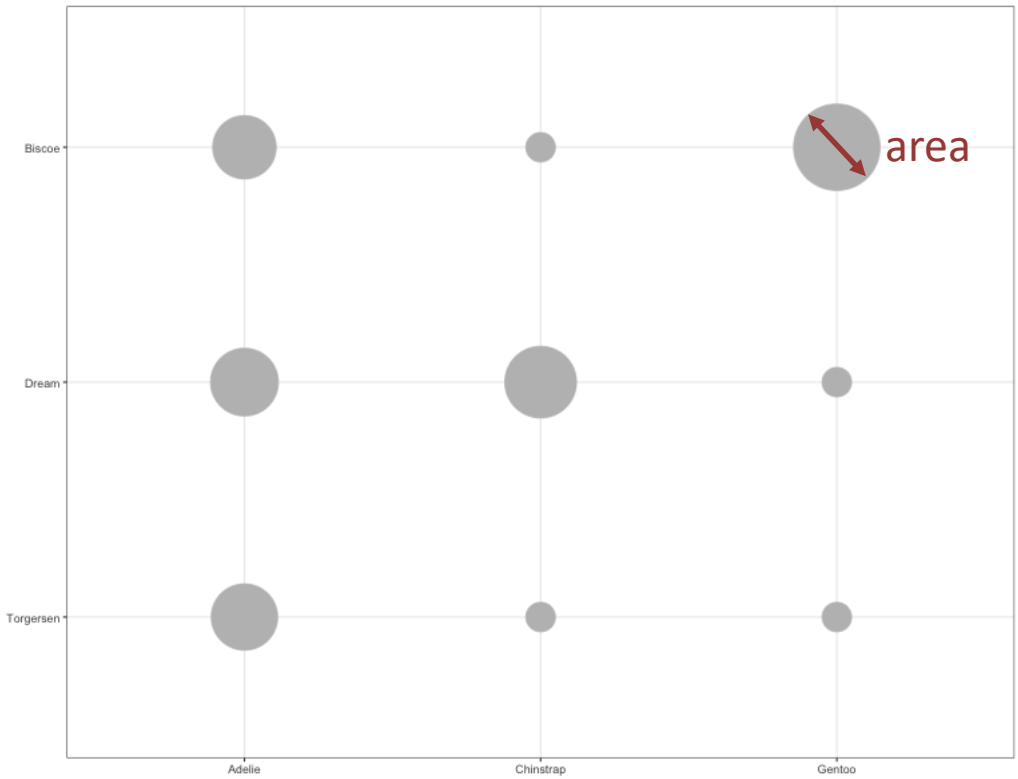


Boxplot showing penguins' body mass index per species.

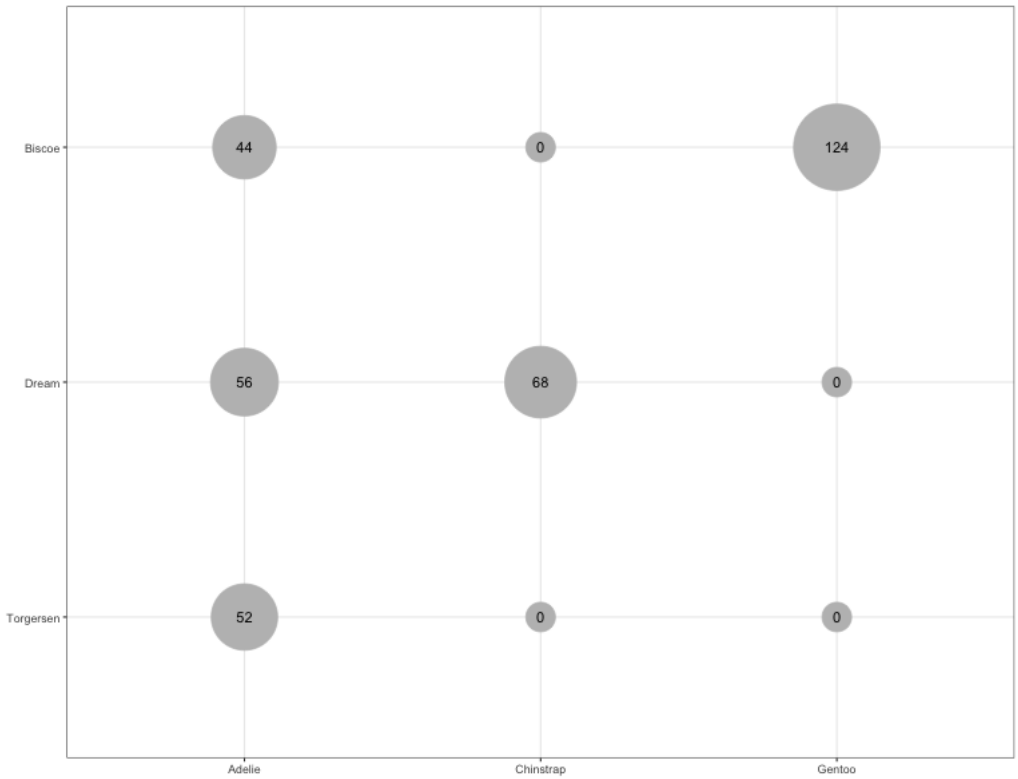


Size comprises different coding possibilities.

Using **SIZE** to convey information.



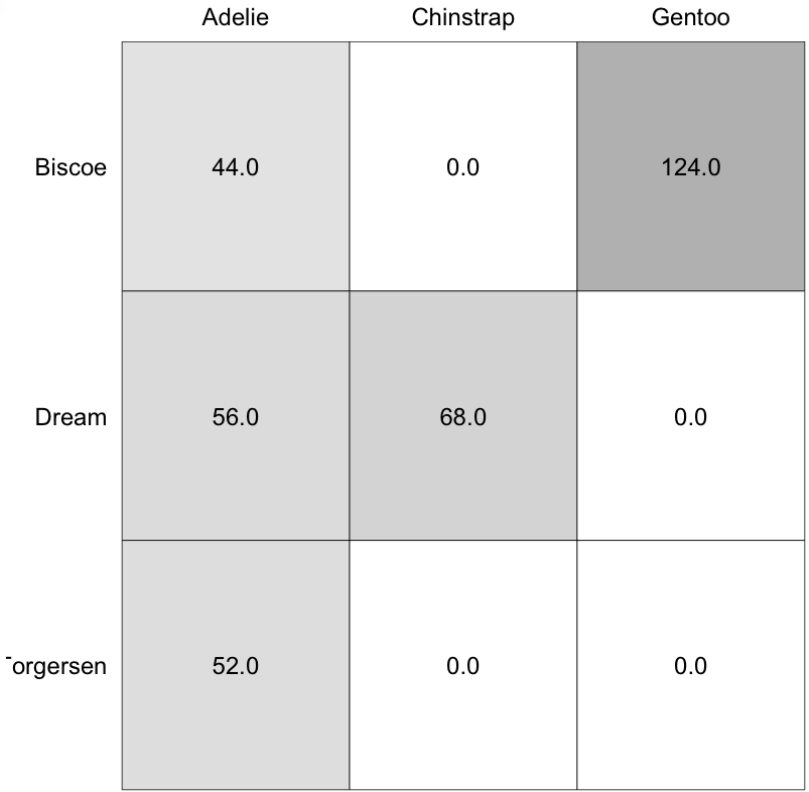
Crosstab table showing number of penguins per species and living space.



Crosstab table showing number of penguins per species and living space.

Be careful using color! See chapter about color perception!

Using **VALUE (SATURATION/BRIGHTNESS)** to convey information.



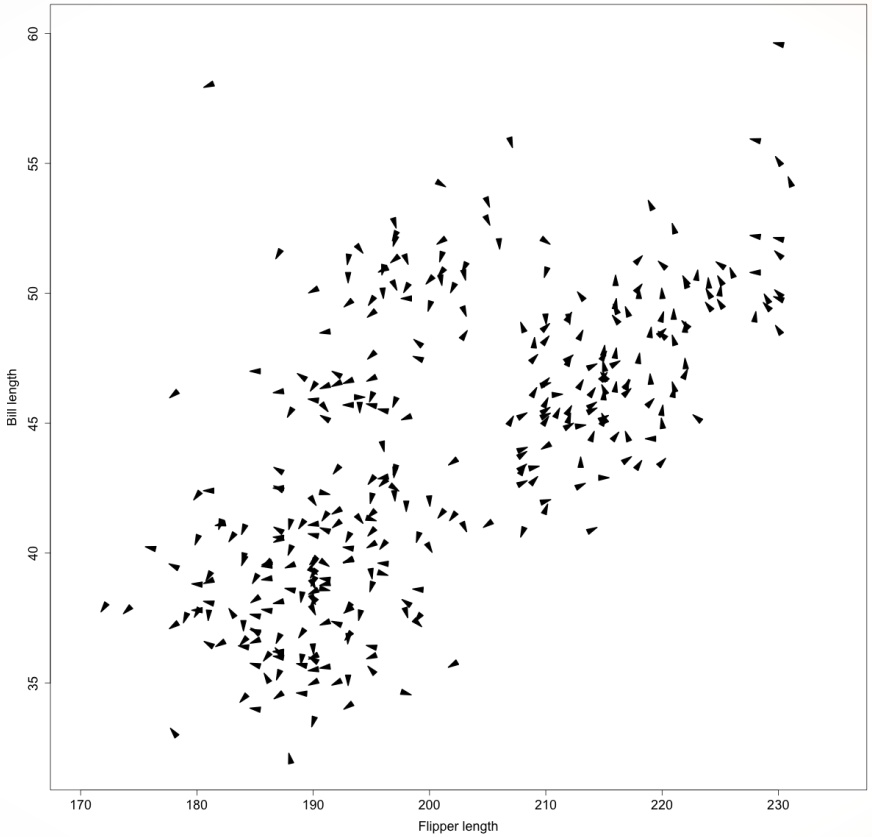
Colored table showing number of penguins per species and living space.

Which rectangle represents the higher value?

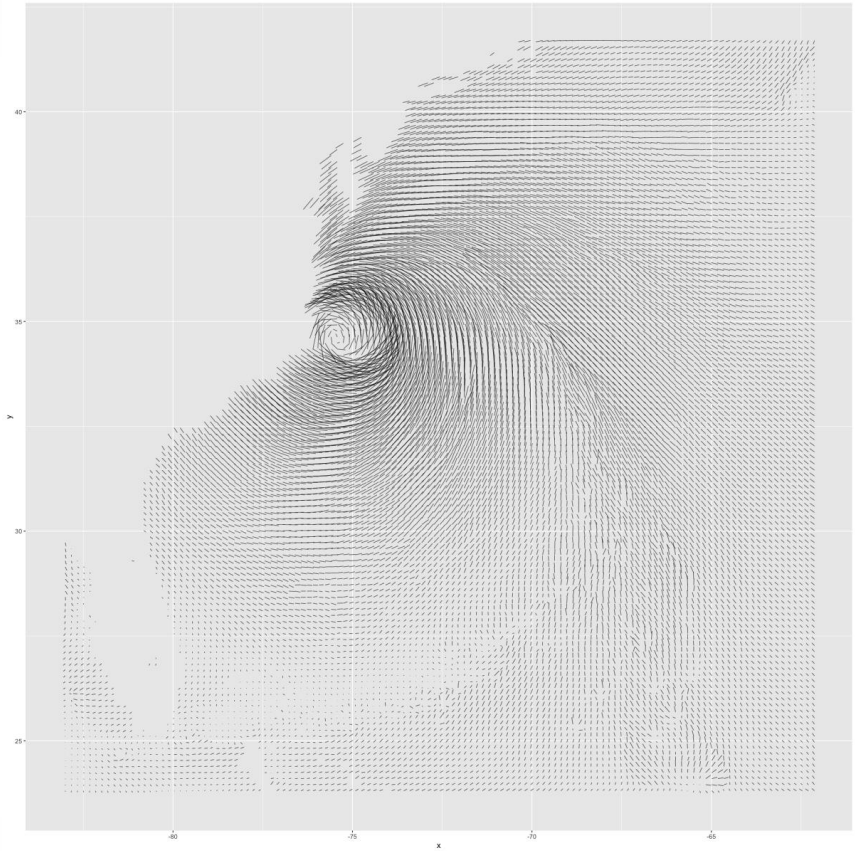
Using **ORIENTATION/ANGLE/DIRECTION** to convey information.

Which direction represents the max or min value?

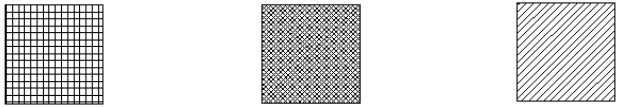
Using **ORIENTATION/ANGLE/DIRECTION** to convey information.



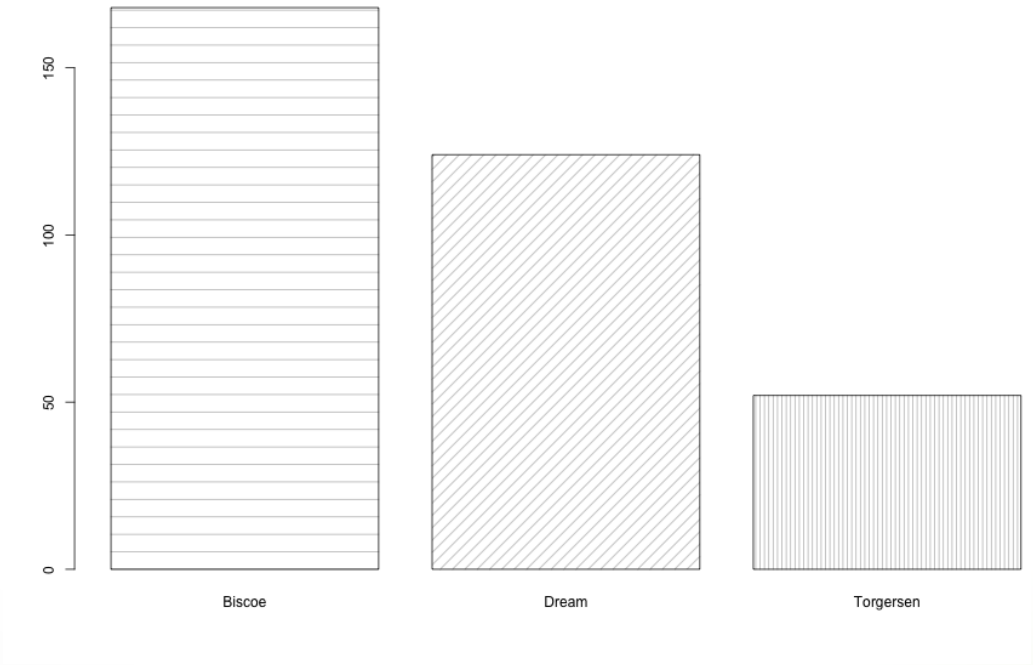
Scatterplot showing penguins' bill depth (orientation), bill length and and flipper length.



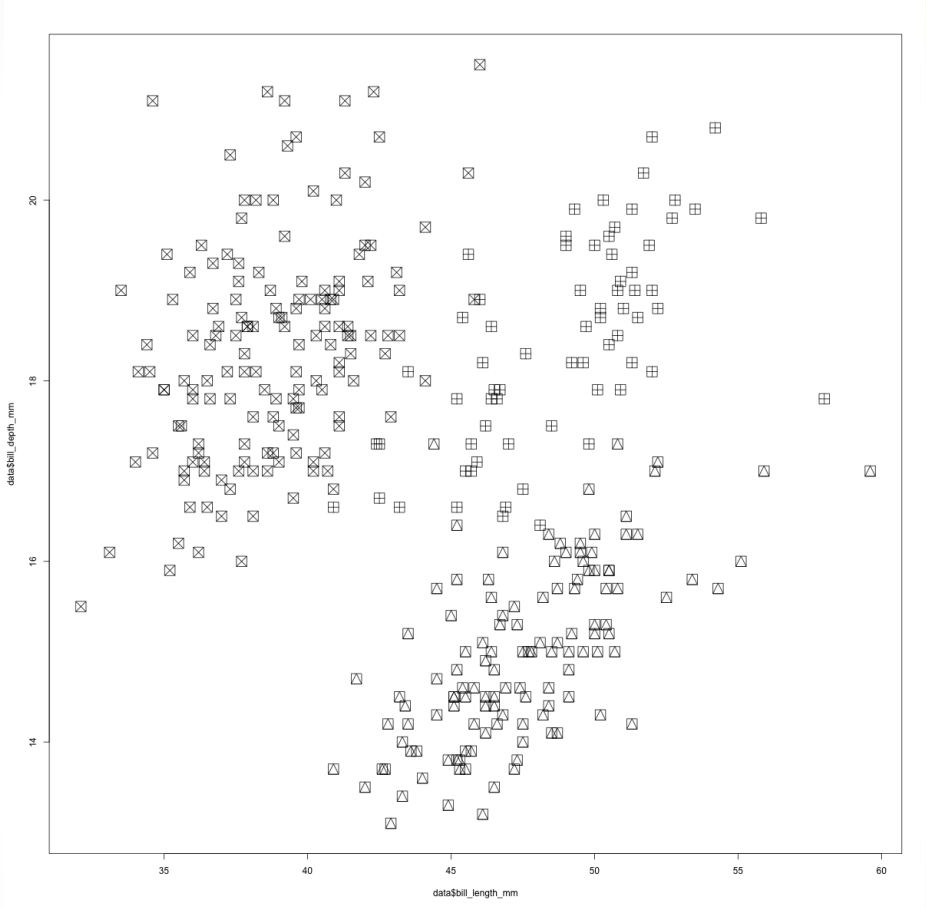
Simulation of hurricane Isabel in 2003.



Using **TEXTURE** to convey information.

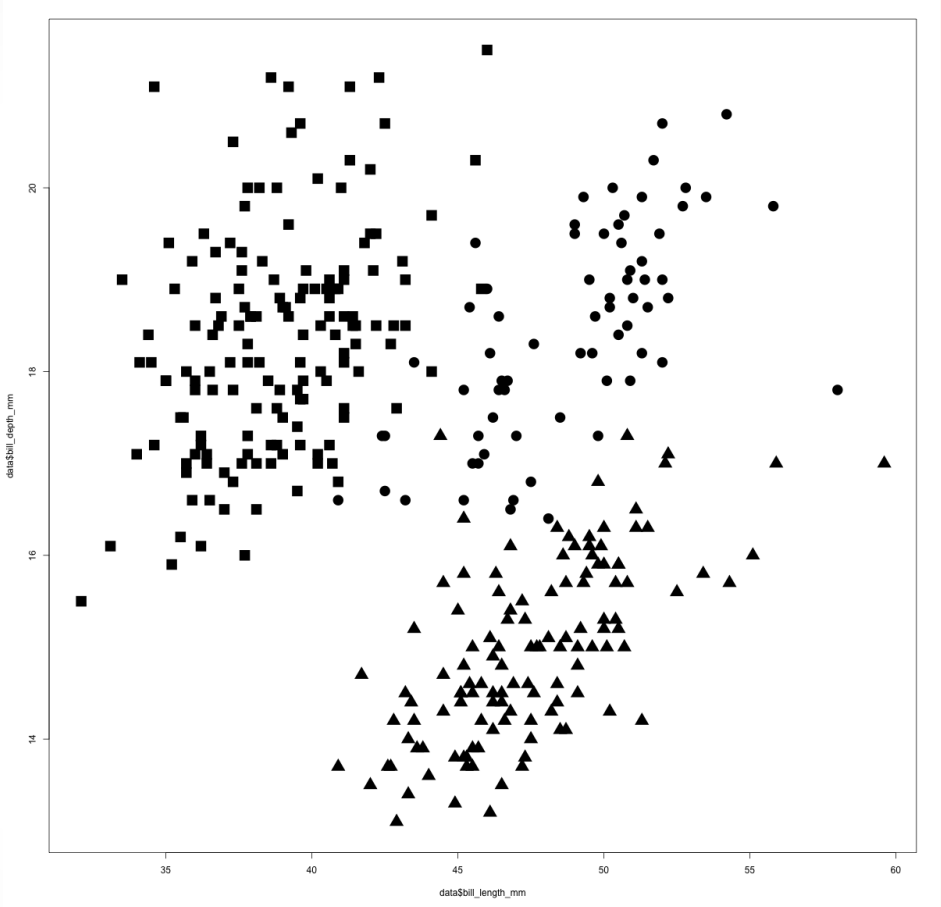


Barchart showing penguins' living space.



Scatterplot showing penguins' species and the respective bill length and bill depth.

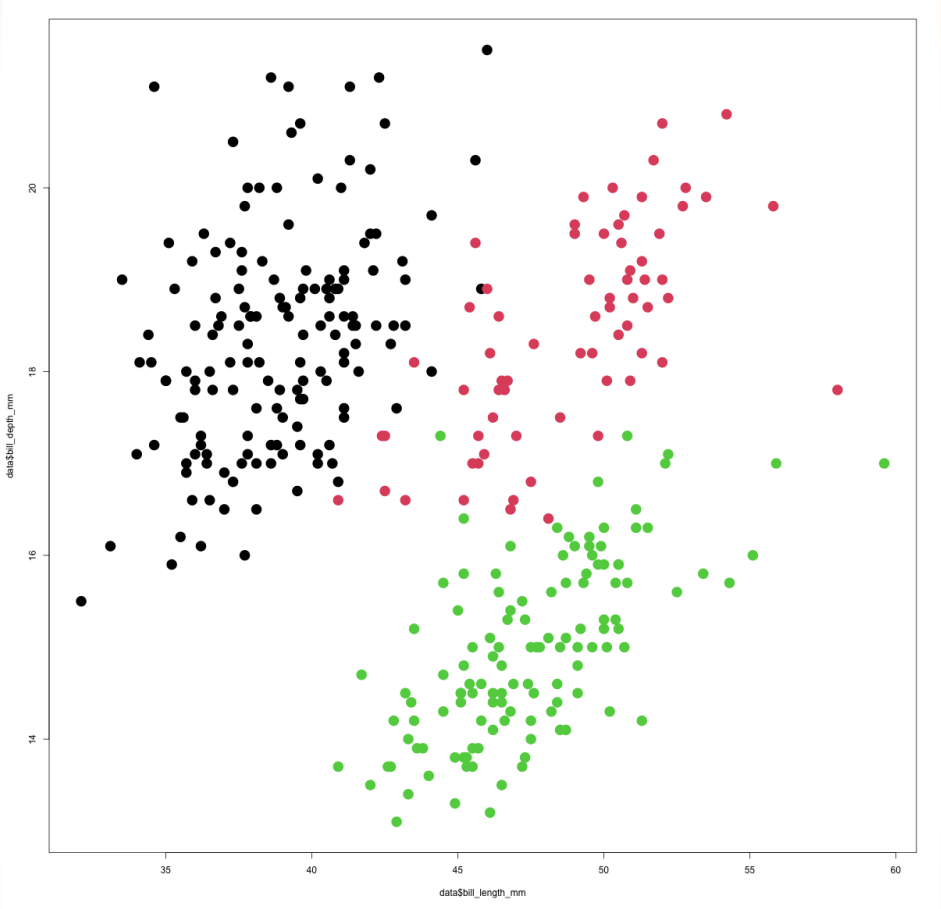
Using **SHAPE** to convey information.



Scatterplot showing penguins' species and the respective bill length and bill depth.

Be careful using color!

Using **COLOR HUE** to convey information.



Scatterplot showing penguins' species and the respective bill length and bill depth.

Summary: Visual Variables

- Position
- Size (length/area/volume)
- Color value (Brightness/Saturation)
- Color hue
- Orientation
- Texture
- Shape
- (Motion)

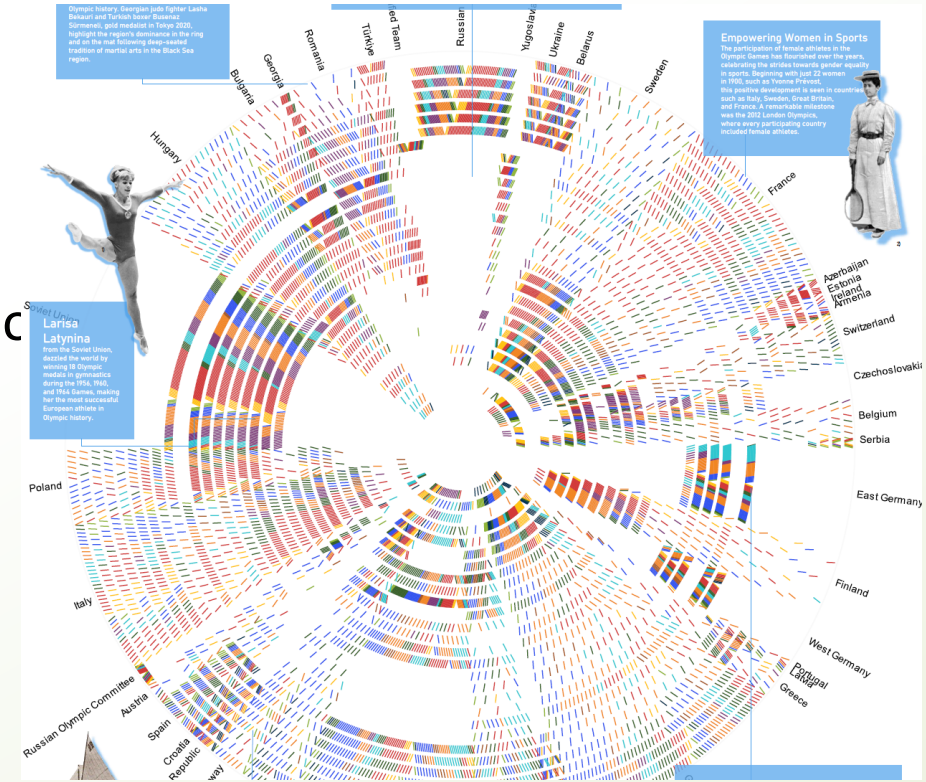
Summary: Visual Variables

- Position
- Size (length/area/volume)
- Color value (Brightness/Saturation)
- Color hue
- Orientation
- Texture
- Shape
- (Motion)

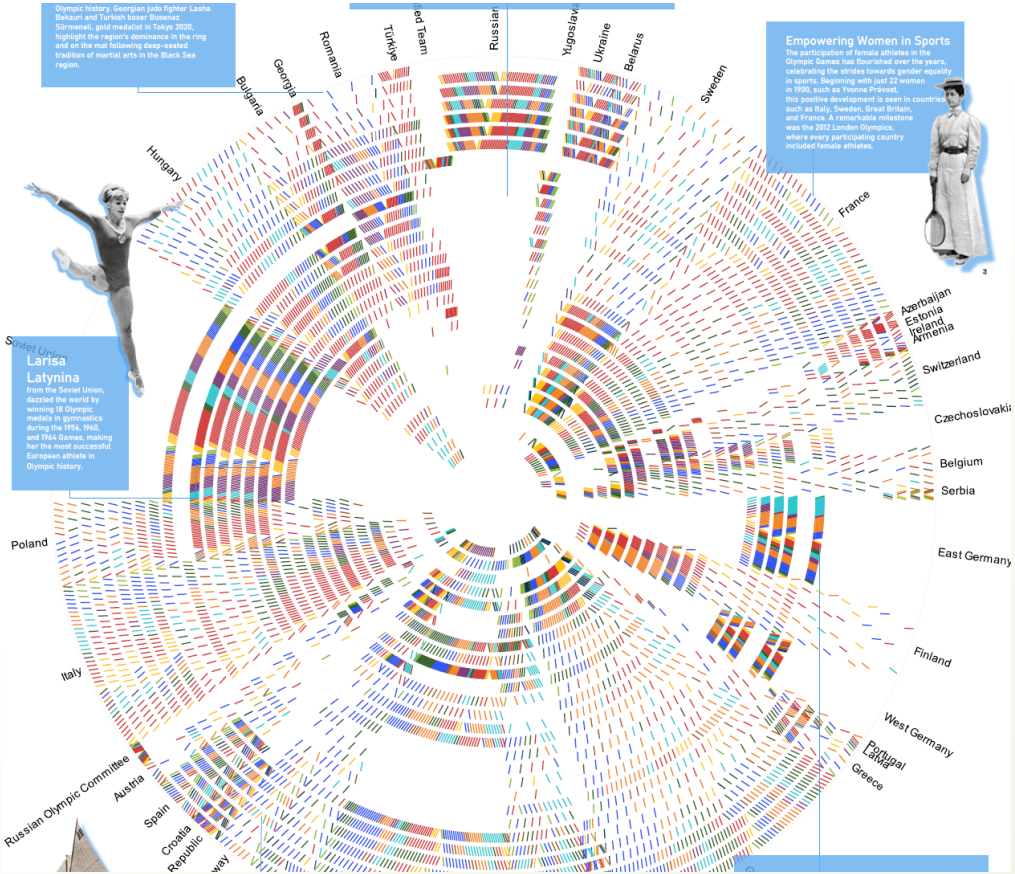
Not exhaustive but a good starting point

Summary: Visual Variables

- Position
- Size (length/area/volume)
- Color value (Brightness/Saturatic
- Color hue
- Orientation
- Texture
- Shape
- (Motion)



Which visual variables are included here?



Summary: Visual Variables

More visual variables:

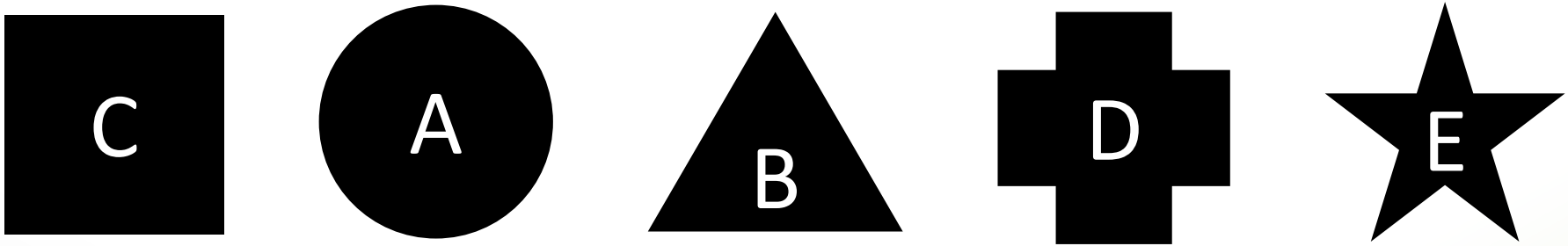
- MacEachren, A. M. (2004). *How maps work: representation, visualization, and design*. Guilford Press.
- Boukhelifa, N., Bezerianos, A., Isenberg, T., & Fekete, J. D. (2012). *Evaluating sketchiness as a visual variable for the depiction of qualitative uncertainty*. *IEEE Transactions on Visualization and Computer Graphics*, *18*(12), 2769-2778.

Summary: Decomposing Visualizations

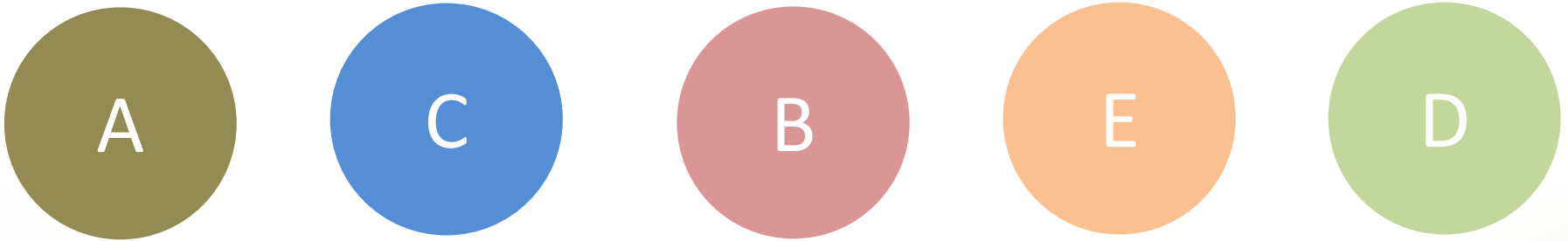
Given the **eight** visual variables, it is possible to decompose most visualizations.

However:

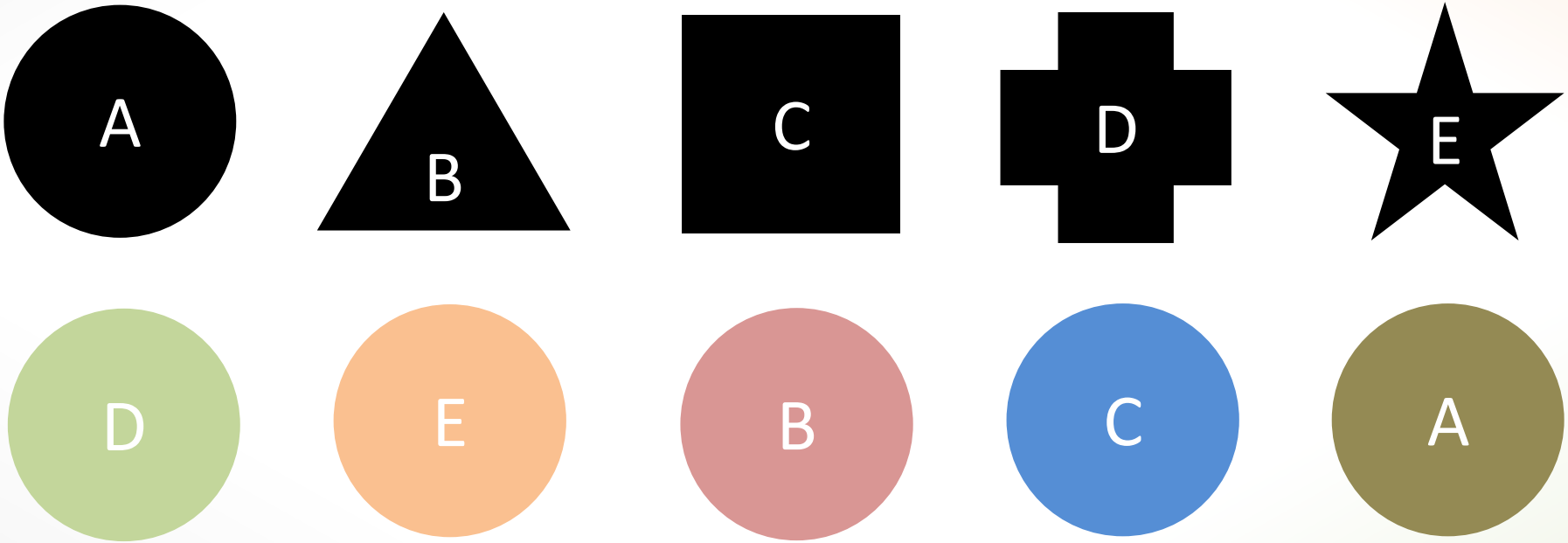
- Are the visual variables appropriately chosen?
- Can we somehow evaluate a visualization?
- How could we evaluate a visualization?



Try to order the **shapes** from low (left) to high (right)



Try to order the **circles** from low (left) to high (right)



Some visual variables are not suitable for displaying **ordinal** or **numeric** data types.

Effects of Visual Variables

Different visual variables can serve different purposes.

- Selective
- Associative
- Ordered
- Quantitative

Effects of Visual Variables

Selective: If a mark changes in this variable and as an effect can be selected from the other marks easily.

- All visual variables (disputable)

Effects of Visual Variables

Associative: All factors have the same visibility. Therefore, a change does not cause the visibility of the signs to vary.

- Texture
- Color hue
- Shape
- Orientation
- Position

Effects of Visual Variables

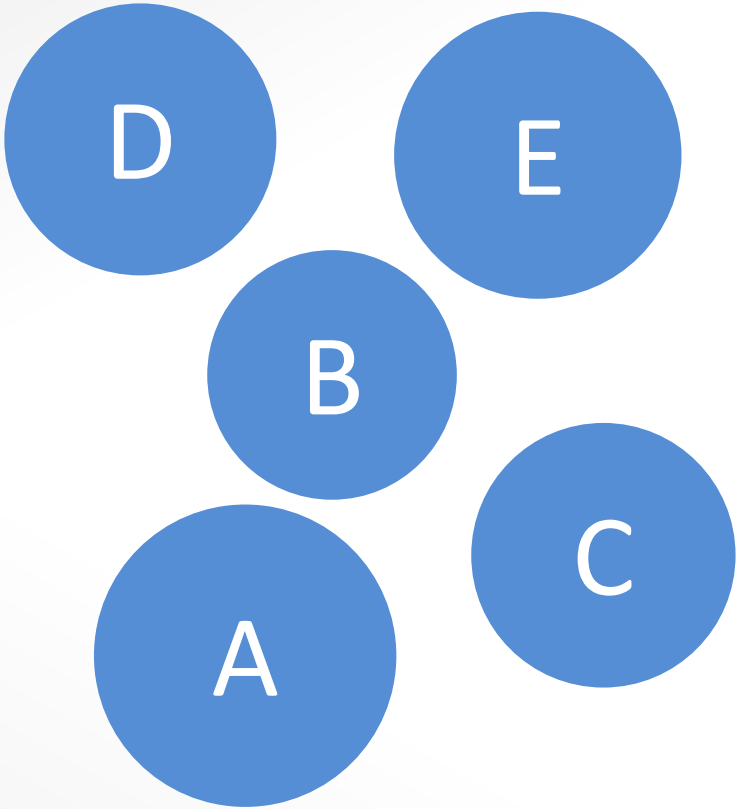
Ordered: Different data values are spontaneously ordered by humans.

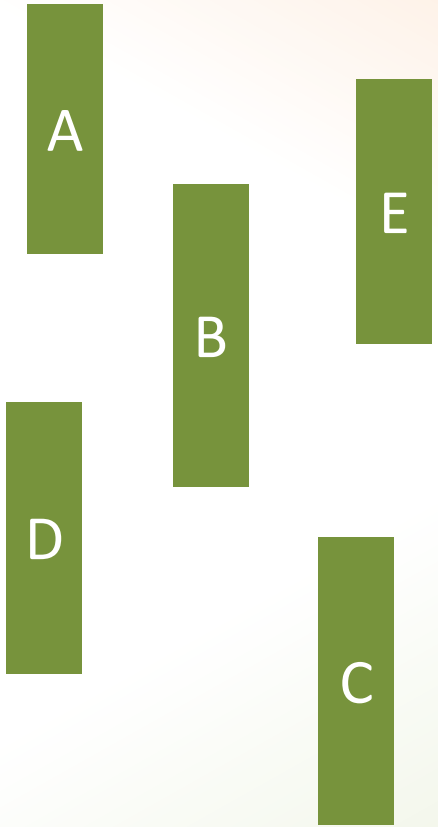
- Size
- Position
- Color value
- Orientation (disputable)

Effects of Visual Variables

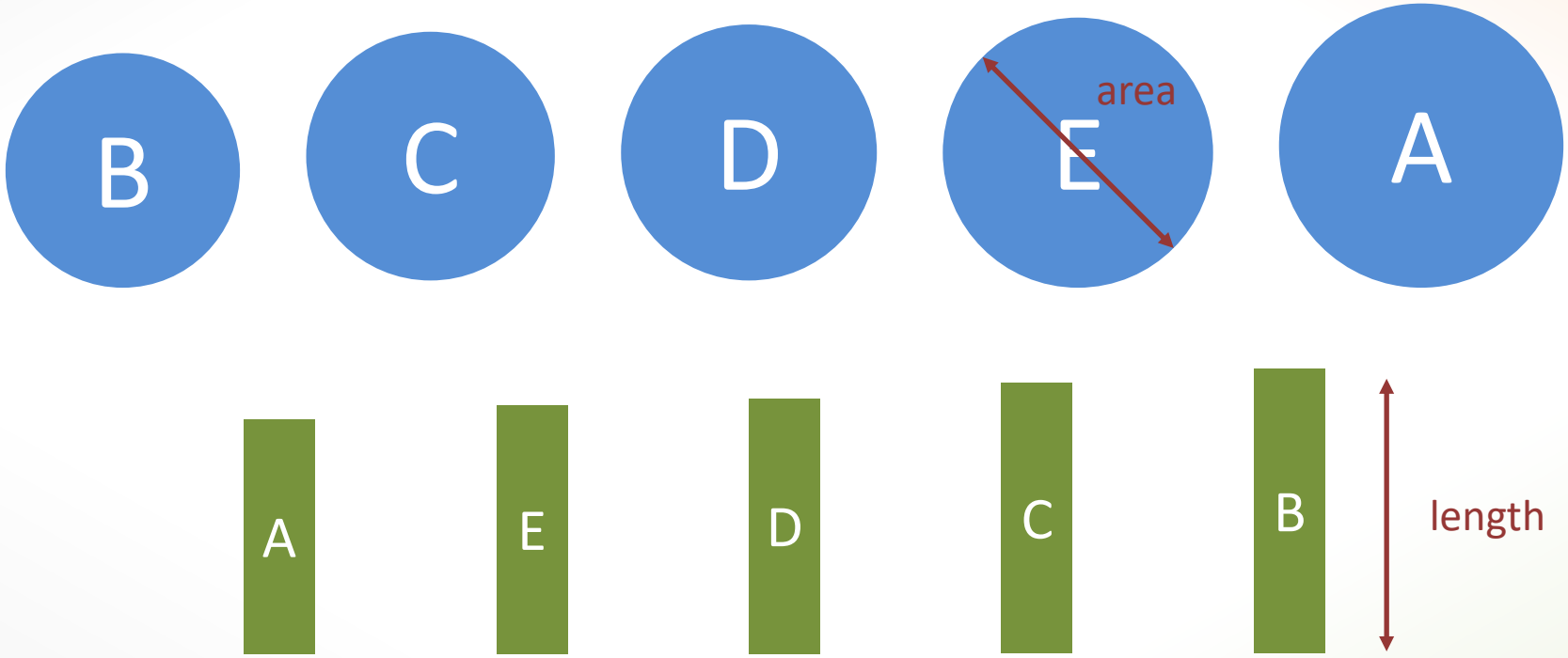
Proportional: These variables obtain a direct association of the relative size.

- Size
- Position
- Color value (disputable)
- Orientation (disputable)





Try to order the **circles** and the **lines** from low (left) to high (right)



There seems to be a difference in performance when considering **size** as visual variable.

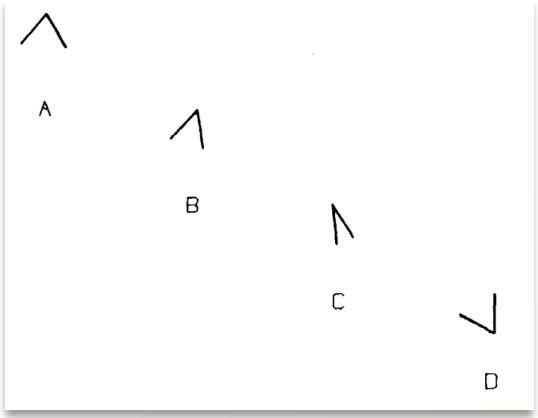
There seems to be a difference in performance when considering **different** visual variables.

Evaluating Visual Variables

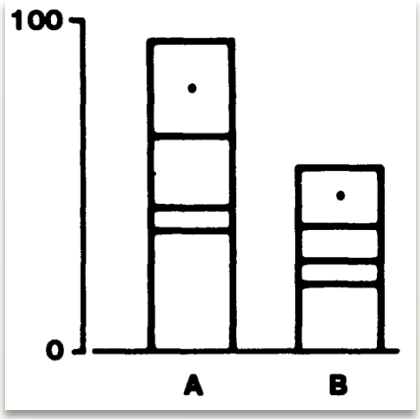
- There seems to be a difference in how well certain visual variables communicate ratios/quantities.
- Several user studies have been conducted to investigate the suitability of visual variables.
- As a result, many suggestions on how to order visual variables exist.
- For simplicity reasons, we just have a look at two different works.
  - Graphical perception – Cleveland and McGill (1984)
  - Crowdsourcing graphical perception – Heer and Bostock (2010)

Ranking of Visual Variables

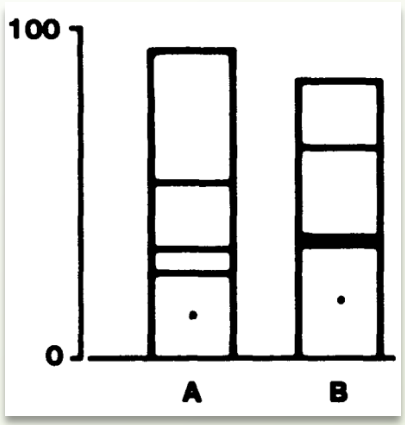
- Participants had to judge the ratio of different visual variables to a given stimulus.



orientation



length



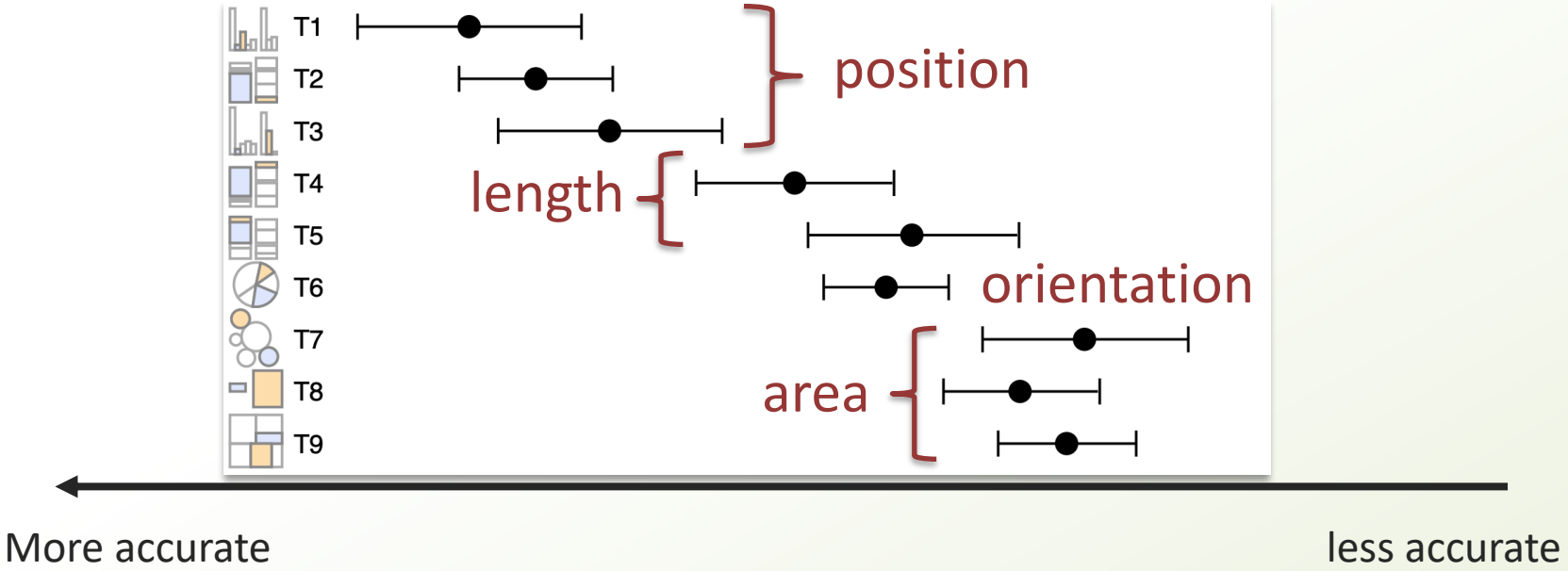
position

Ranking of Visual Variables

- There was a difference in error rate.
- Summary after different experiments:

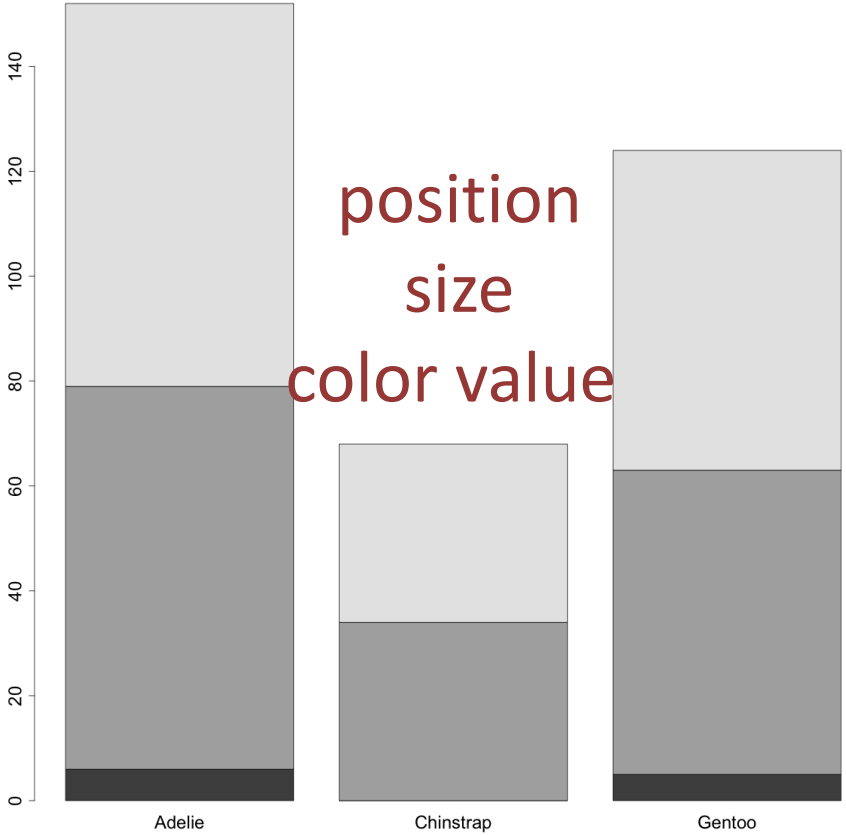
Ranking of Visual Variables

- Replicated the previous study as a crowdsourced experiment:

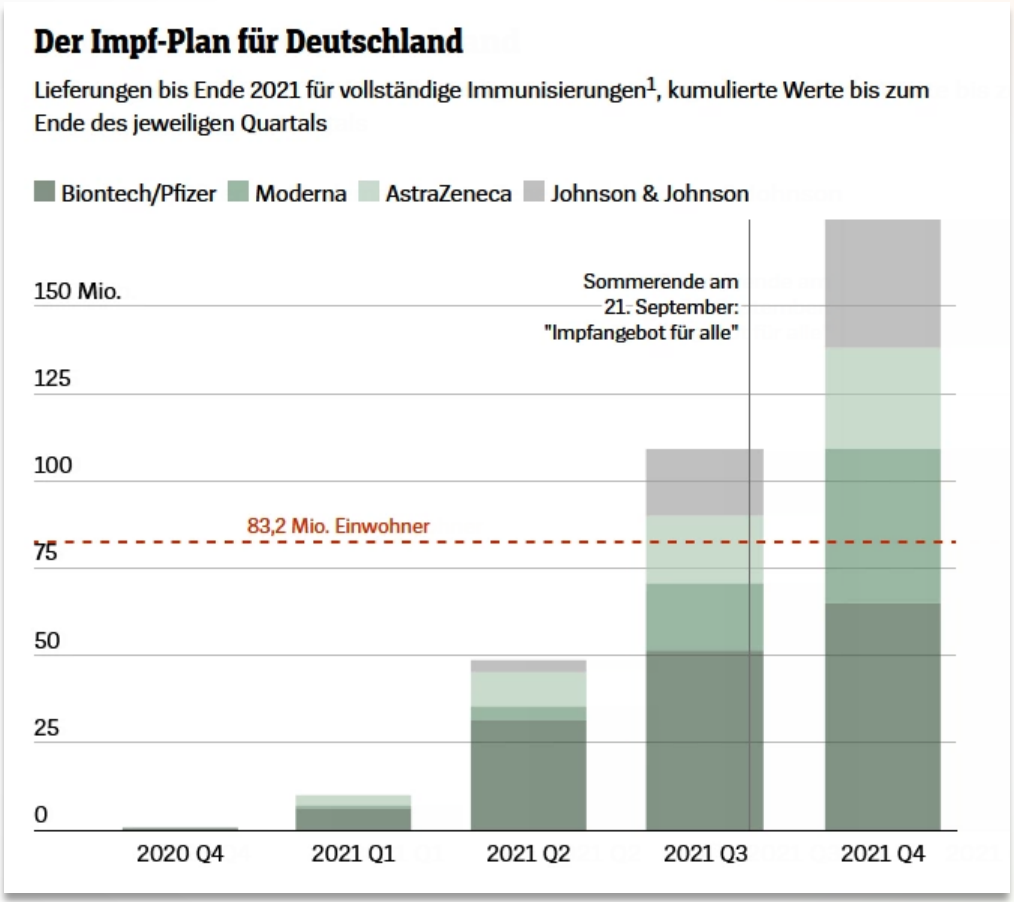


Summary: Ranking of Visual Variables

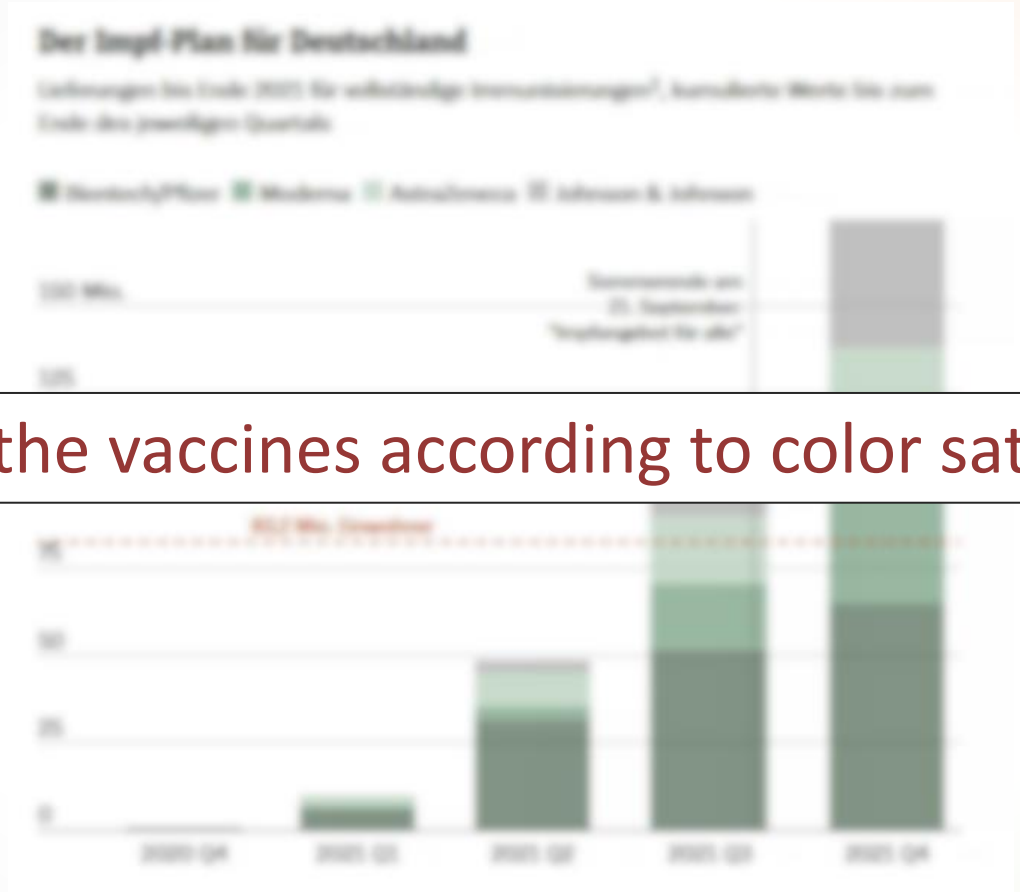
- Visual variables can serve different purposes.
- There is a difference in performance when judging visual variables.
- Can we make use of this knowledge to evaluate a visualization?



Are the visual variables appropriate to represent the underlying data?

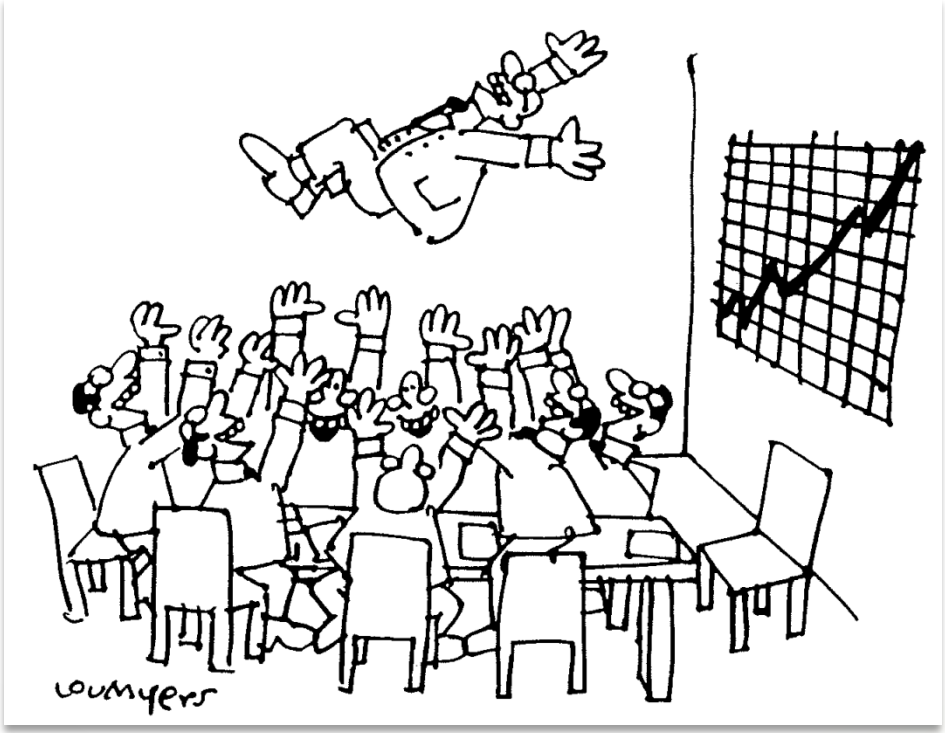


Supply of different vaccines per quarter.



Ranking the vaccines according to color saturation.

Supply of different vaccines per quarter.



Why is the hero rewarded?

New Yorker cartoon from April 4, 1983.

Evaluating a Visualization

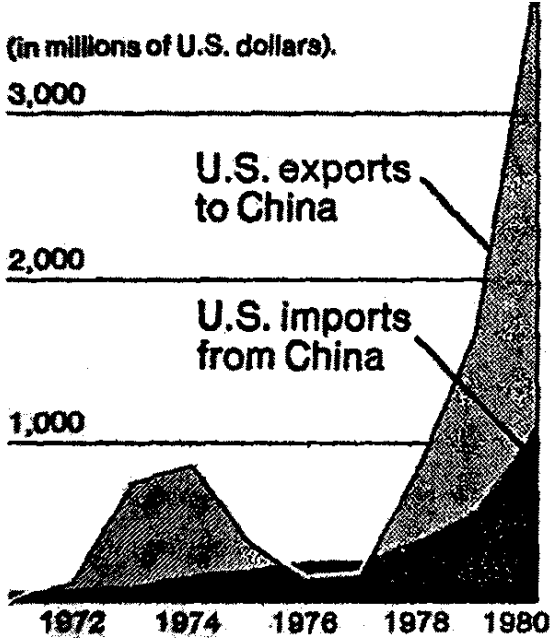
We are focusing on two different questions:

1. **Effectiveness**: Is there an alternative representation, which would be better (participants are more accurate/faster in decoding the information)?
2. **Expressiveness**: Is the visualization communicating the information – and only the information

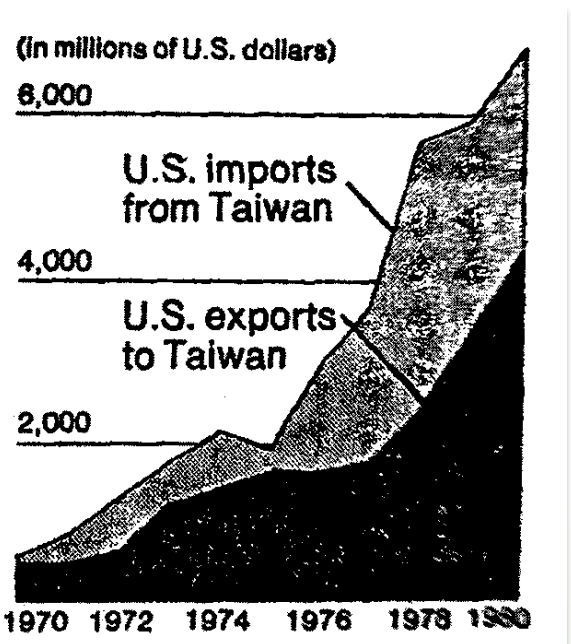
Violating the Effectiveness

- Poor choice of visual variables.
  - Encode the most important information in the most effective way (principle of importance ordering).
- Inconsistent mapping.
  - Do not change the data encoding during the presentation of results.

U.S. trade with China



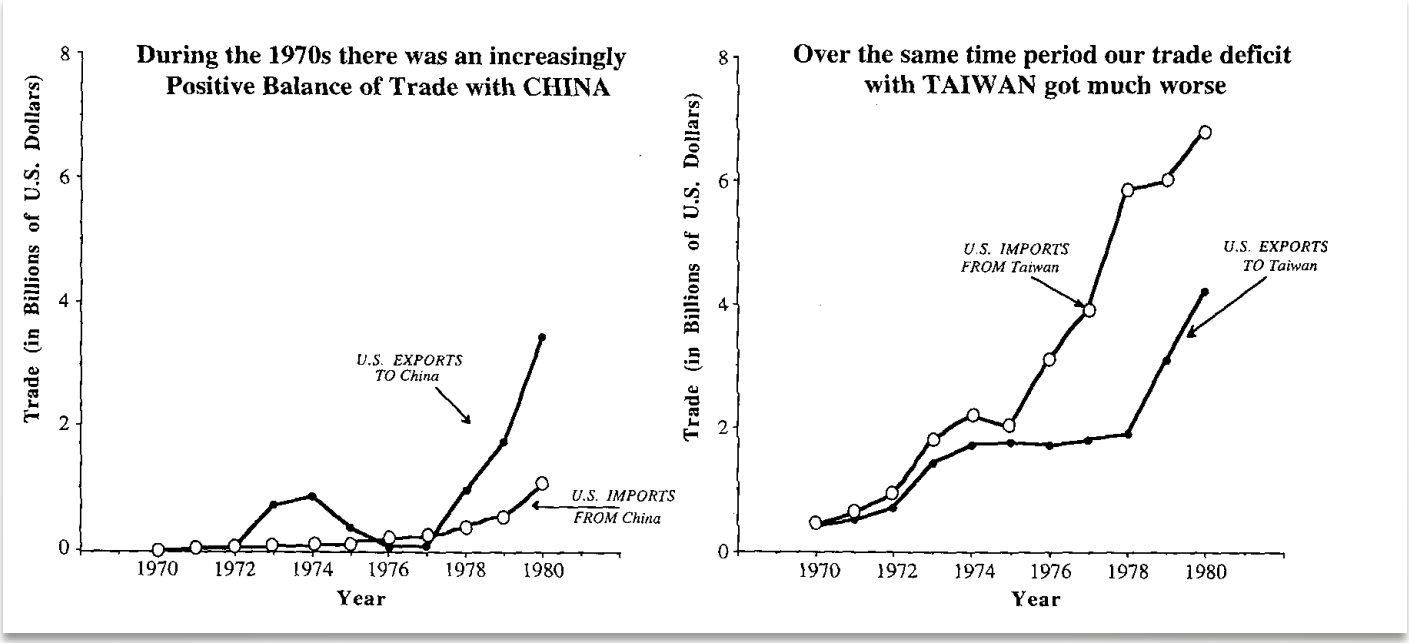
U.S. trade with Taiwan



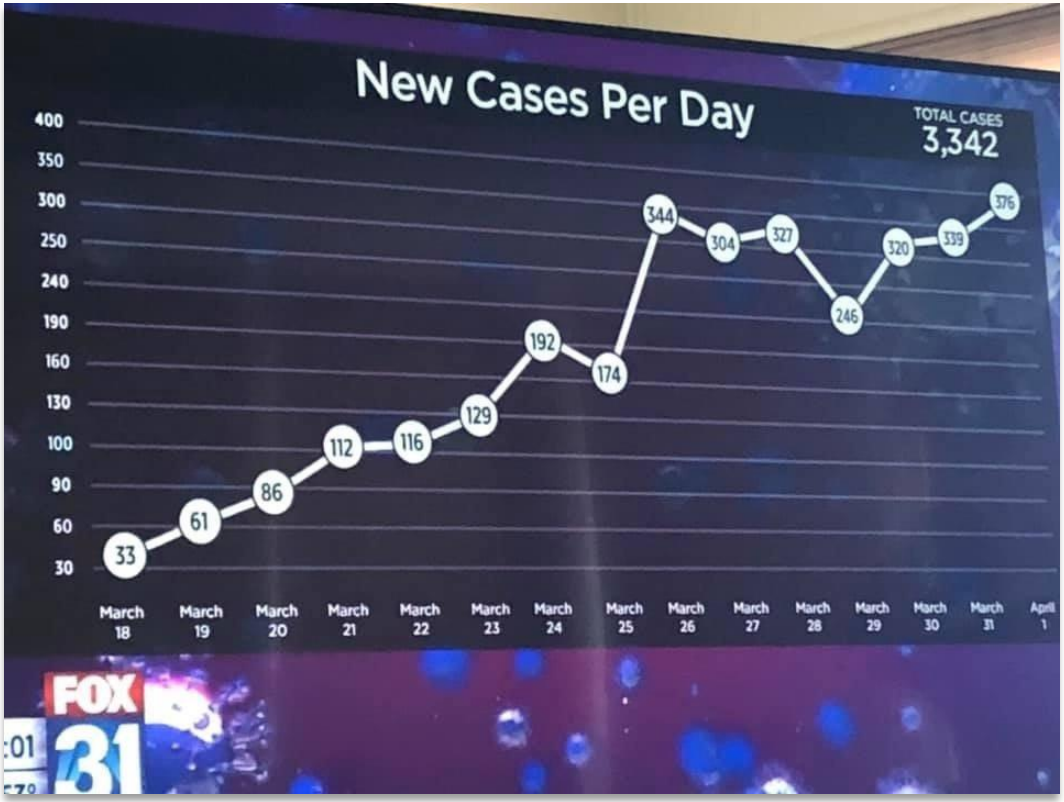
Investigating U.S. trade: Comparison of exports and imports with China and Taiwan.

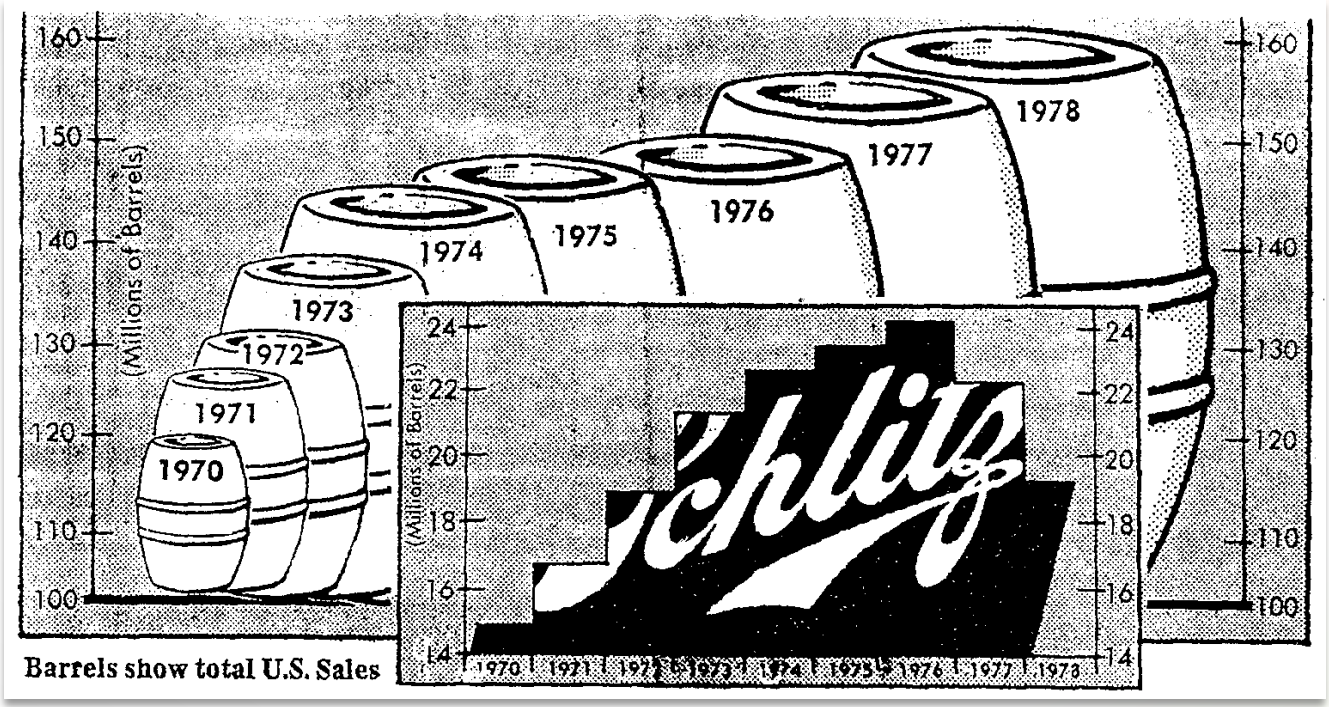
U.S. trade with China

U.S. trade with Taiwan

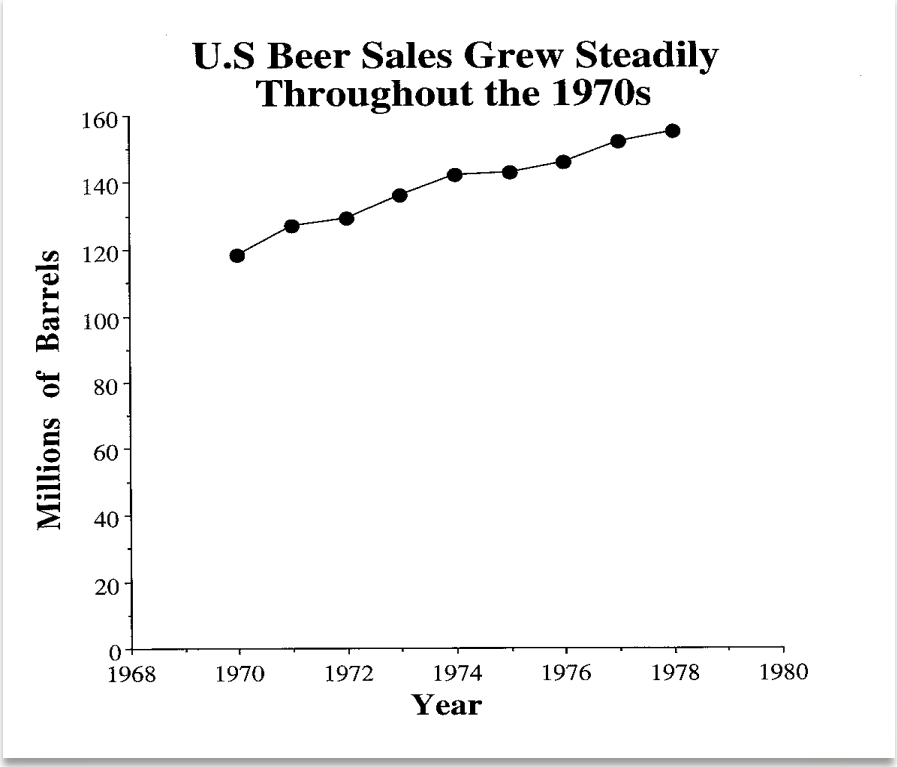


Reworked the visualization: Identical scales and consistent mapping of imports and exports.

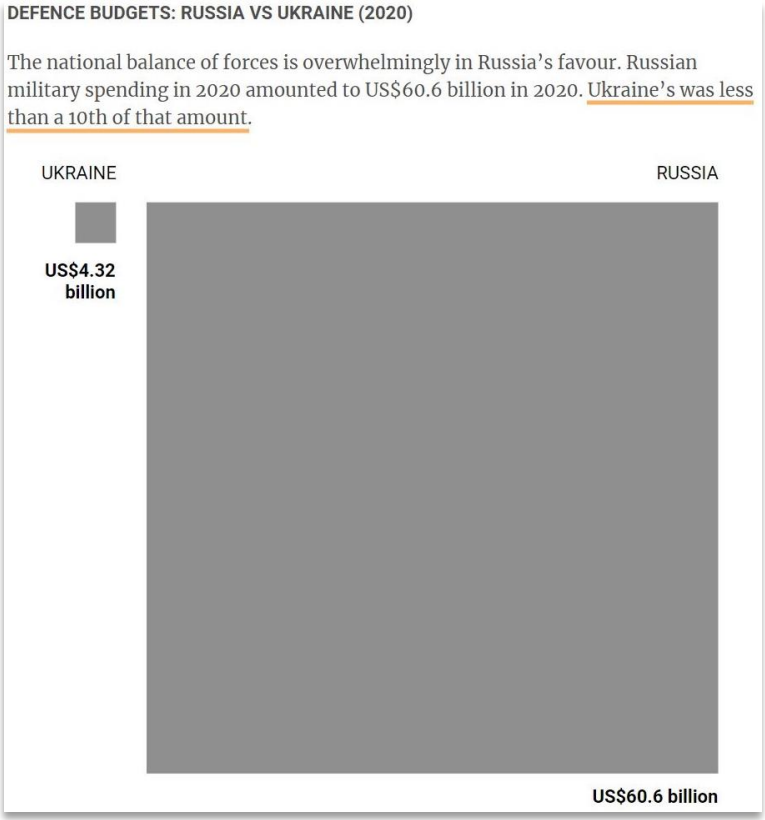

New Cases Per Day
TOTAL CASES
3,342
400
350
300
250
240
190
160
130
100
90
60
30
33
61
86
112
116
129
192
174
344
304
327
246
320
339
376
March 18
March 19
March 20
March 21
March 22
March 23
March 24
March 25
March 26
March 27
March 28
March 29
March 30
March 31
April 1
FOX
31

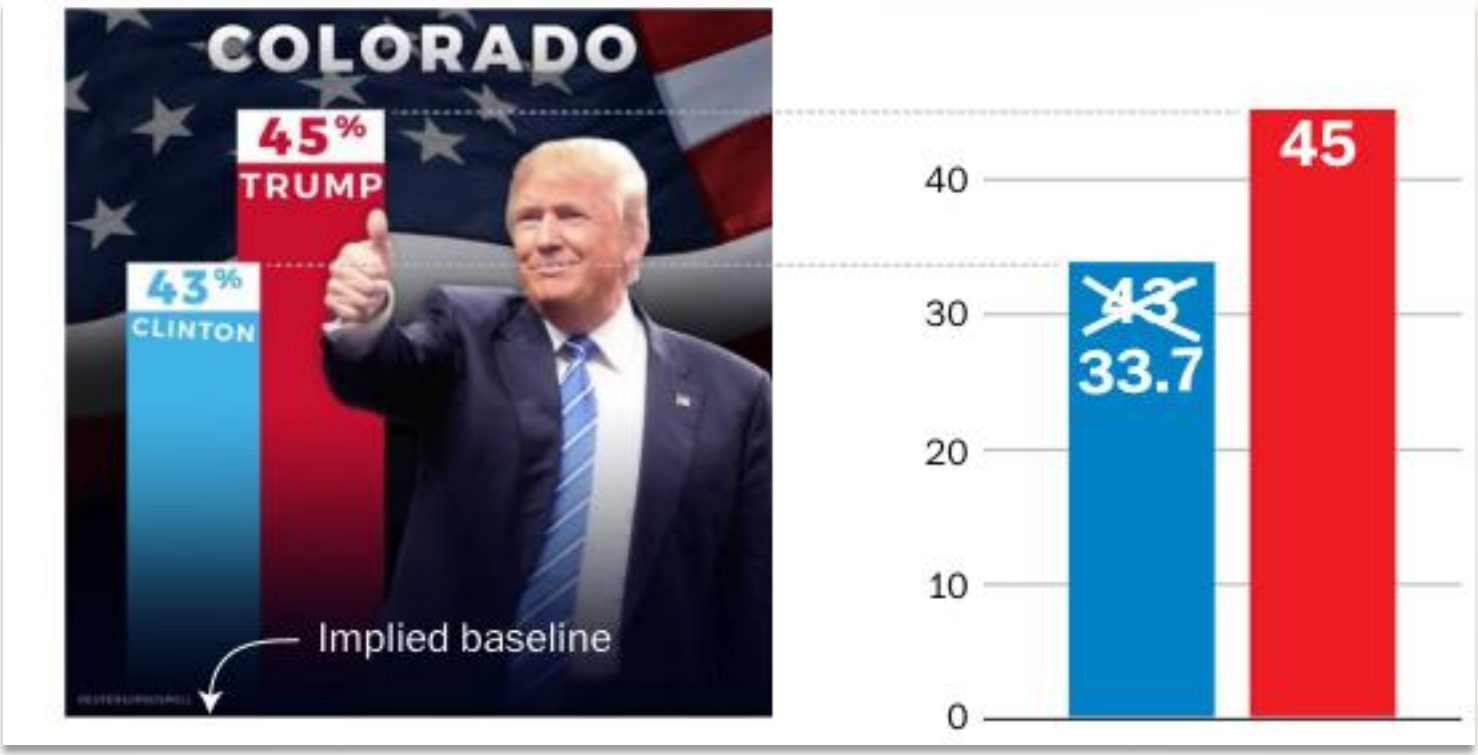


U.S. Sales & Market Share of Schlitz company.

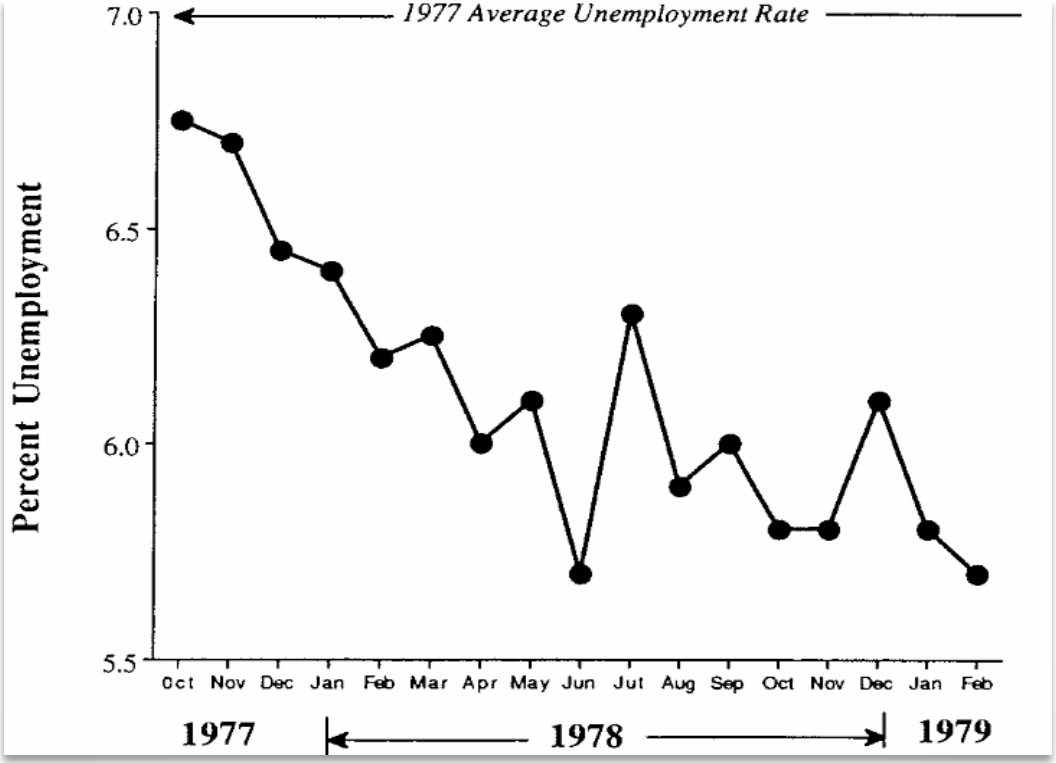


Reworked: Removing the barrels (volume) and changing the scale.

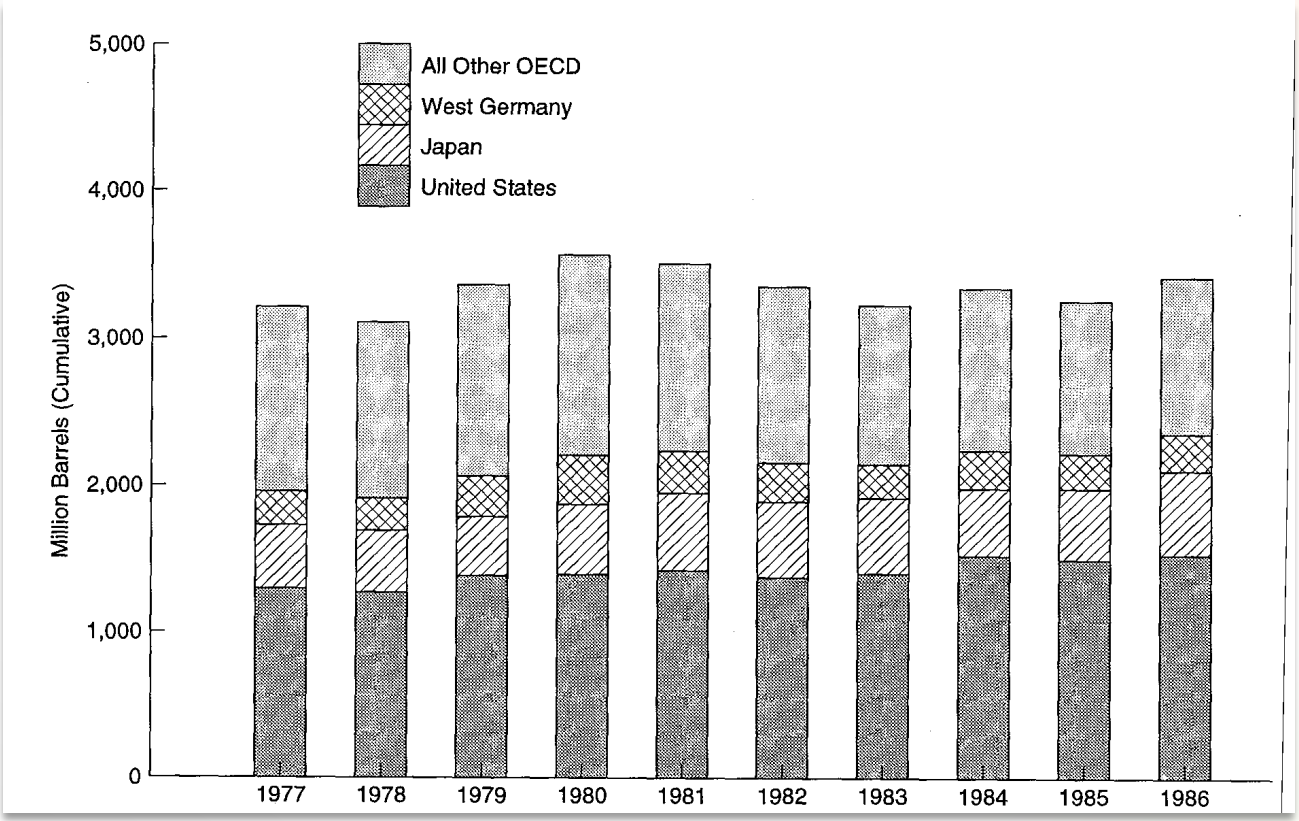

DEFENCE BUDGETS: RUSSIA VS UKRAINE (2020)
The national balance of forces is overwhelmingly in Russia's favour. Russian military spending in 2020 amounted to US$60.6 billion in 2020. Ukraine's was less than a 10th of that amount.
UKRAINE
RUSSIA
US$4.32 billion
US$60.6 billion


COLORADO
45%
TRUMP
43%
CLINTON
Implied baseline
40
30
20
10
0
43
33.7
45

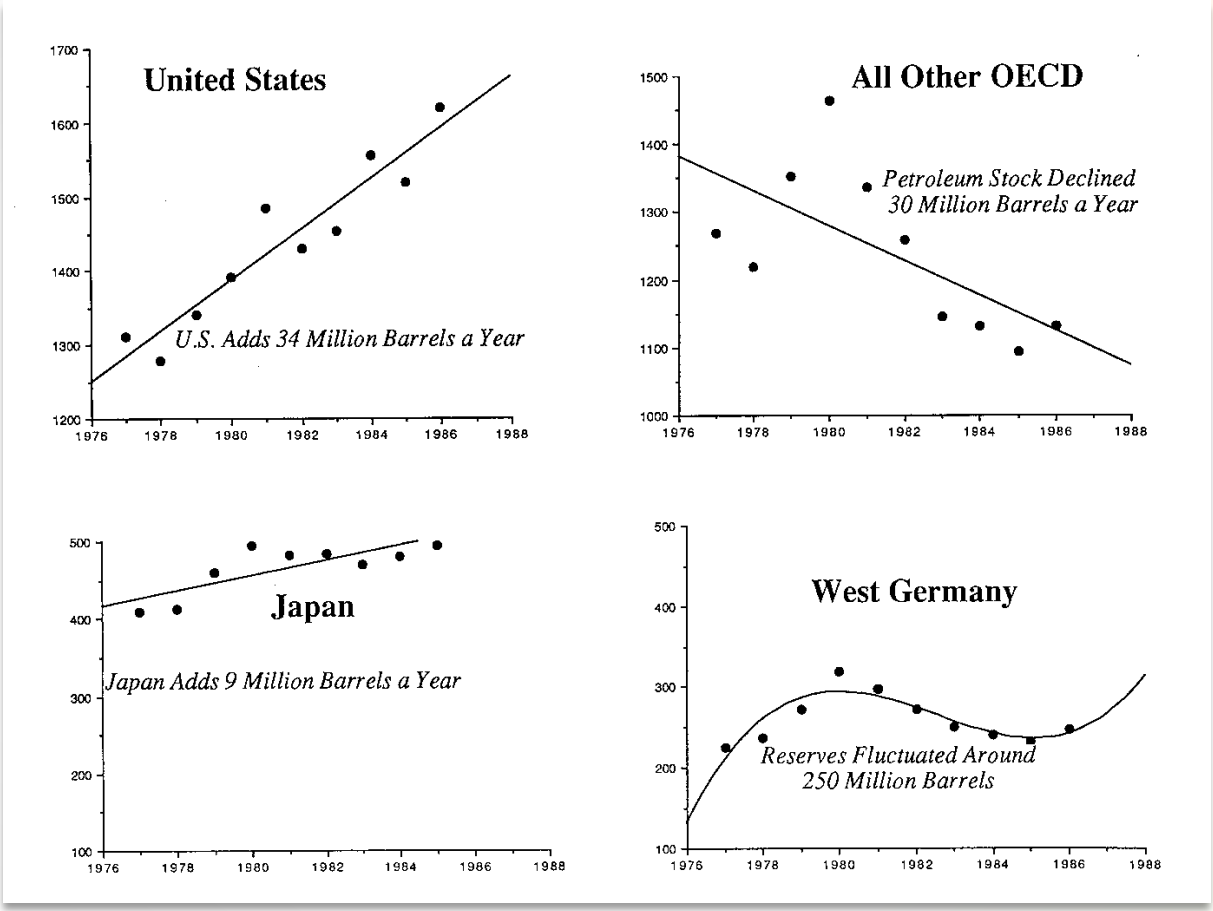
Unemployment rate in the year 1978 and 1979.



Reworked: Introduced a new starting point (Oct 1977) and extended the scale to see smaller changes.



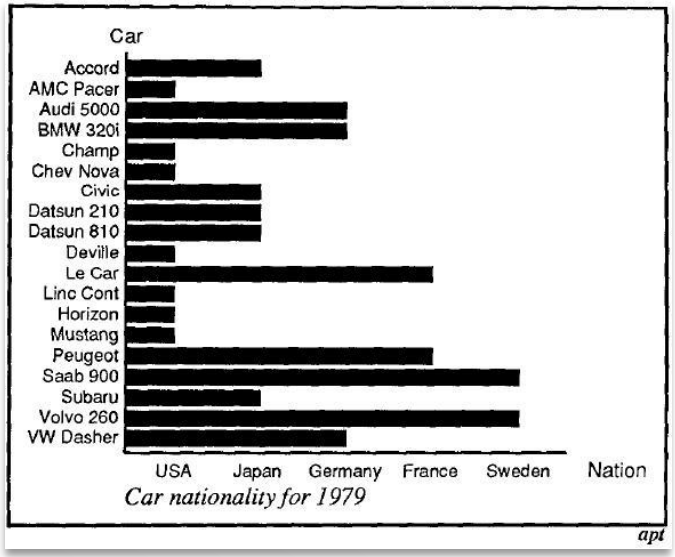
Showing the changes in primary stocks of petroleum in OECD countries

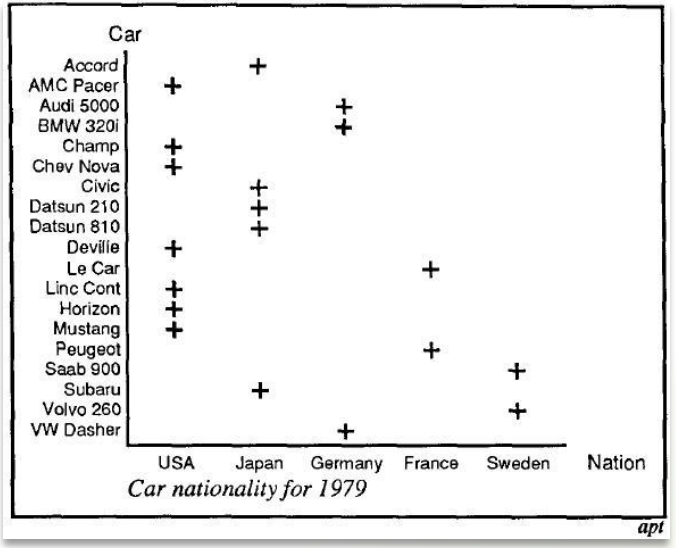


Reworked: separated the individual time series. Added a model (line) to highlight trends.

Violating the Expressiveness

- Introducing information that is not in the data.
  - Carefully chose the visual variables. Think about their purpose (associative, ordered, …)



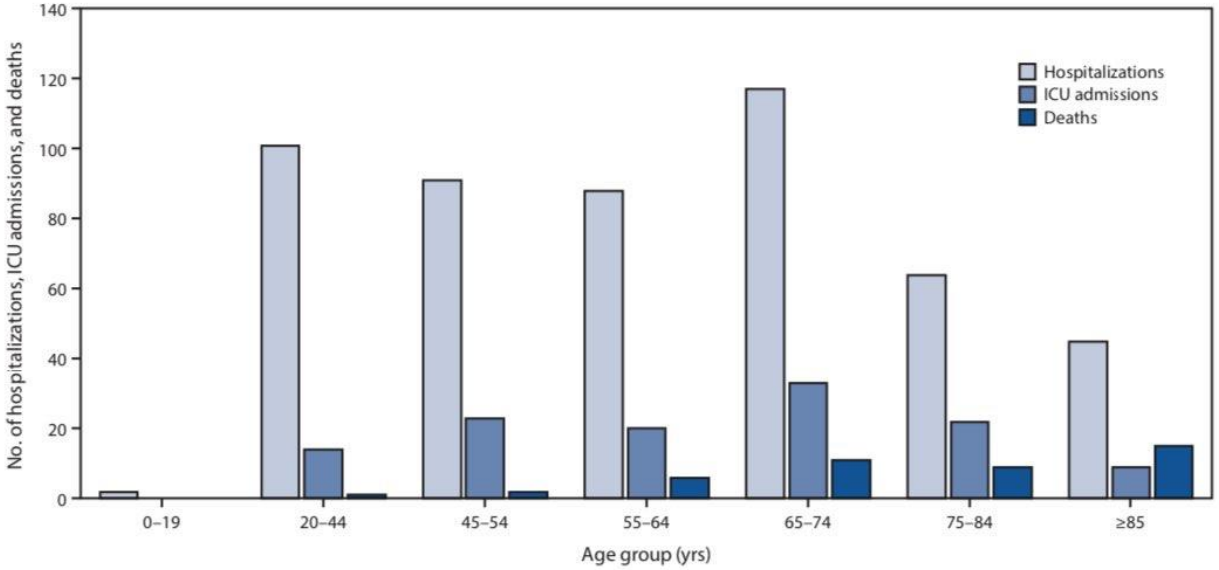


Visualization of cars and the manufacturing country.

	USA	Japan	Germany	France	Sweden
AMC Pacer	X				
Audi 5000			X		
BMW 320			X		
Champ	X				
Chev Nova	X				
Civic		X			
Le Car				X	
Accord		X			

Visualization of cars and the manufacturing country.

FIGURE 2. COVID-19 hospitalizations,* intensive care unit (ICU) admissions,† and deaths,§ by age group — United States, February 12–March 16, 2020



* Hospitalization status missing or unknown for 1,514 cases.
† ICU status missing or unknown for 2,253 cases.
§ Illness outcome or death missing or unknown for 2,001 cases.

Why is the expressiveness violated?

Take-Home Message

Tell the truth and nothing but the truth

(don't lie, and don't lie by omission)

Use encodings that people decode better

(where better = faster and/or more accurate)

Visualizations are not about "making pretty pictures"

Further Reading/Thinking

Is a chart just a combination of individual visual variables?

- Robert Kosara (2022). *More Than Meets the Eye: A Closer Look at Encodings in Visualization*. CG&A.

Do we really perceive the visual variable as intended by the creator (pie chart)?

- Skau, Drew and Kosara, Robert (2016). *Arcs, angles, or areas: Individual data encodings in pie and donut charts.* Computer Graphics Forum.