Data visualization

Visual perception and encoding

David Hoksza
http://siret.ms.mff.cuni.cz/hoksza

Course outline

- Visual perception, design principles
- Tables and charts visualization
- Web-based visualization, D3.js

Design and programming

- Dimension reduction
- Visualization of networks

Algorithms

- (Infographics)
- Project presentations

"Soft skills"

Course organization

Labs

- Once per two weeks
  - Practical visualization
    - Tableau, D3.js, …
  - Homework for each of the topics – at least two of them need to be submitted
    - Each additional homework will be worth 5 extra points for the exam
- Projects
  - Topic proposal by the end of March
  - Presentation at the end of the semester

Examination

- Written or oral exam (depends on the epidemiological situation)

Classification

- Examination (90%)
- Questions about lectures (10%)

Projects

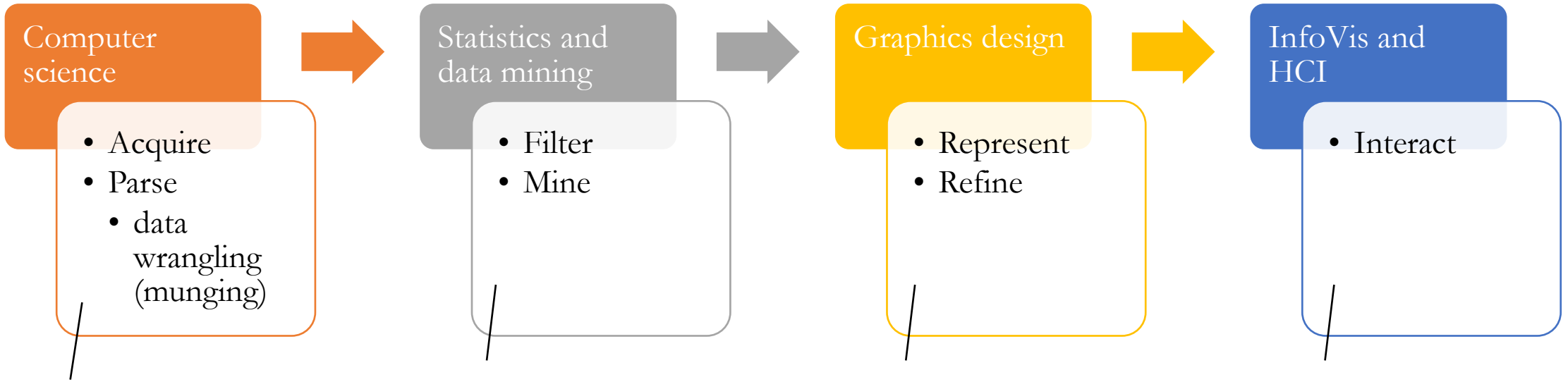
- Slides for presentation

- Problem walkthrough
- Data format description
- Code description
  - Components used
  - Problematic points
  - Interesting points
- Conclusion

Lecture outline

- Motivation
- Visual perception
- Visual encoding
- Data types and relations
- Measures to summarize sets of data

Data science process



Basically ETL (Extraction, Transformation, Loading) known from data warehousing

Core of data science – Exploratory Data Analysis (EDA)

Visual encoding of the numerical data

Ways to let user interact with the data to get new insights

Goals of data visualization

- To **communicate information** clearly and effectively **through graphical means**
  - Data usually has a **structure** which needs to be **revealed** using data visualization → **explore patterns** in the data
  - To help **find** the desired **information** more **effectively** and **intuitively**
    - Picking up things with the naked eye that would otherwise be hidden
- Turning numbers into **story** → **storytelling with data**

Exploratory vs explanatory visualization

You and data

Exploratory visualization

- **What the data is**, what is **hidden** in the data
- Enables users to **look at the data** from different **angles**

Data and your audience

Explanatory visualization

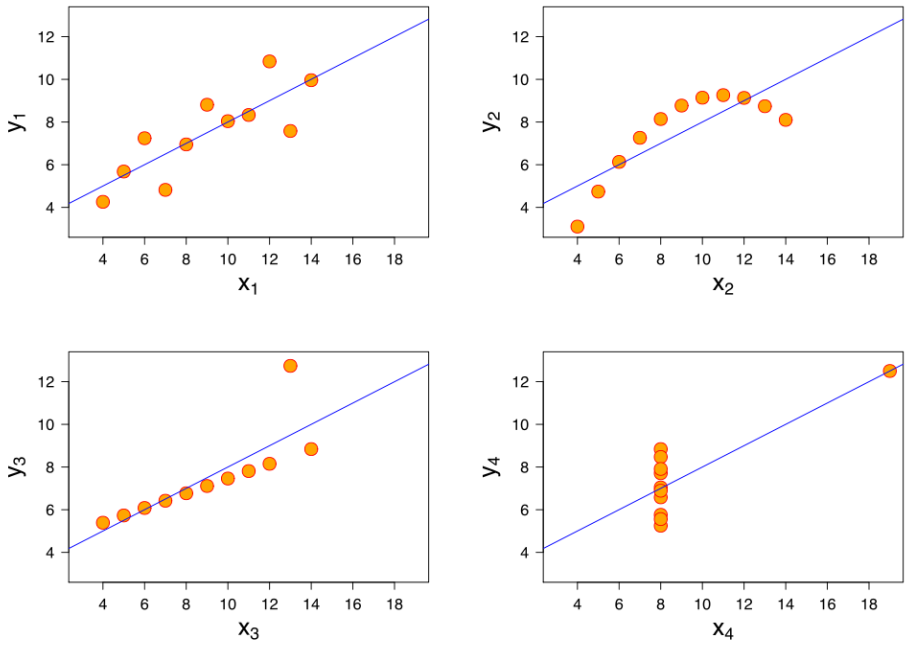
- Helping a user to **make sense of the data** by choosing the right visualization techniques
- Need to know the context from which the audience come and what they need to know
- Strategic **placements of elements** and **choice of attributes** to communicate the information clearly and help users to focus on what is important

Graphics over statistics

- Visualization can reveal/distinguish data/trends/patterns, … which statistics can not

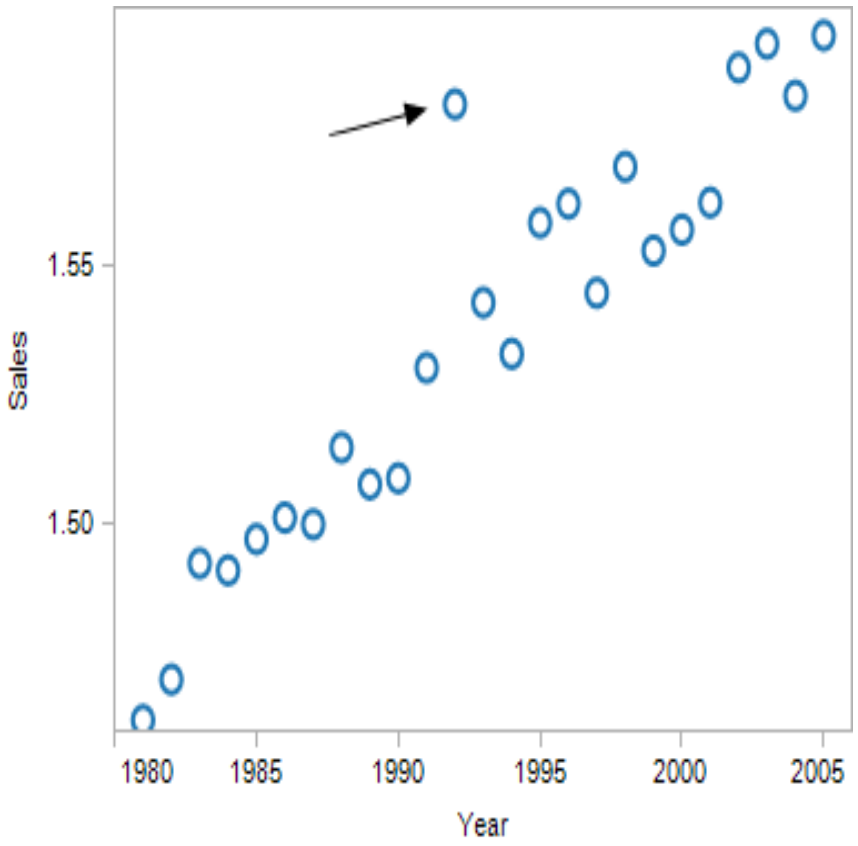
Anscombe's quartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89



Four data sets with nearly identical linear model (mean, variance, linear regression line, correlation coefficient)

	A	B
1	Year	Sales
2	1981	1.4622
3	1982	1.47004
4	1983	1.49253
5	1984	1.49118
6	1985	1.49722
7	1986	1.50138
8	1987	1.50008
9	1988	1.51493
10	1989	1.50781
11	1990	1.50899
12	1991	1.53037
13	1992	1.58137
14	1993	1.54299
15	1994	1.53307
16	1995	1.55845
17	1996	1.56213
18	1997	1.54488
19	1998	1.56927
20	1999	1.55305
21	2000	1.5571
22	2001	1.56235
23	2002	1.58847
24	2003	1.59309
25	2004	1.58303
26	2005	1.5947



Find an outlier….

Visual perception

70%

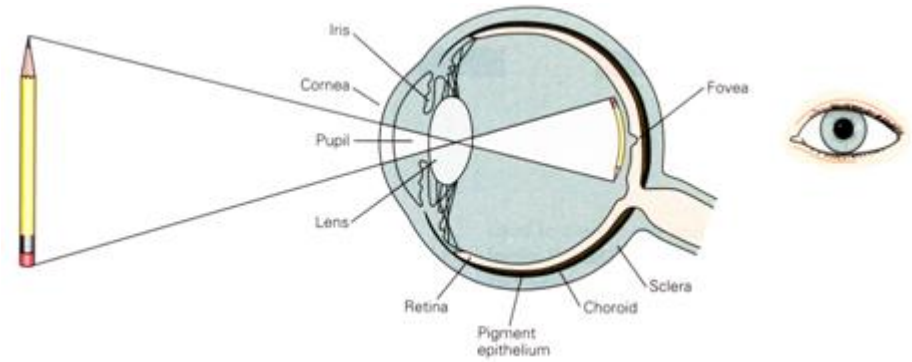
30%

Mechanics of sight

Sensation
(physical process)

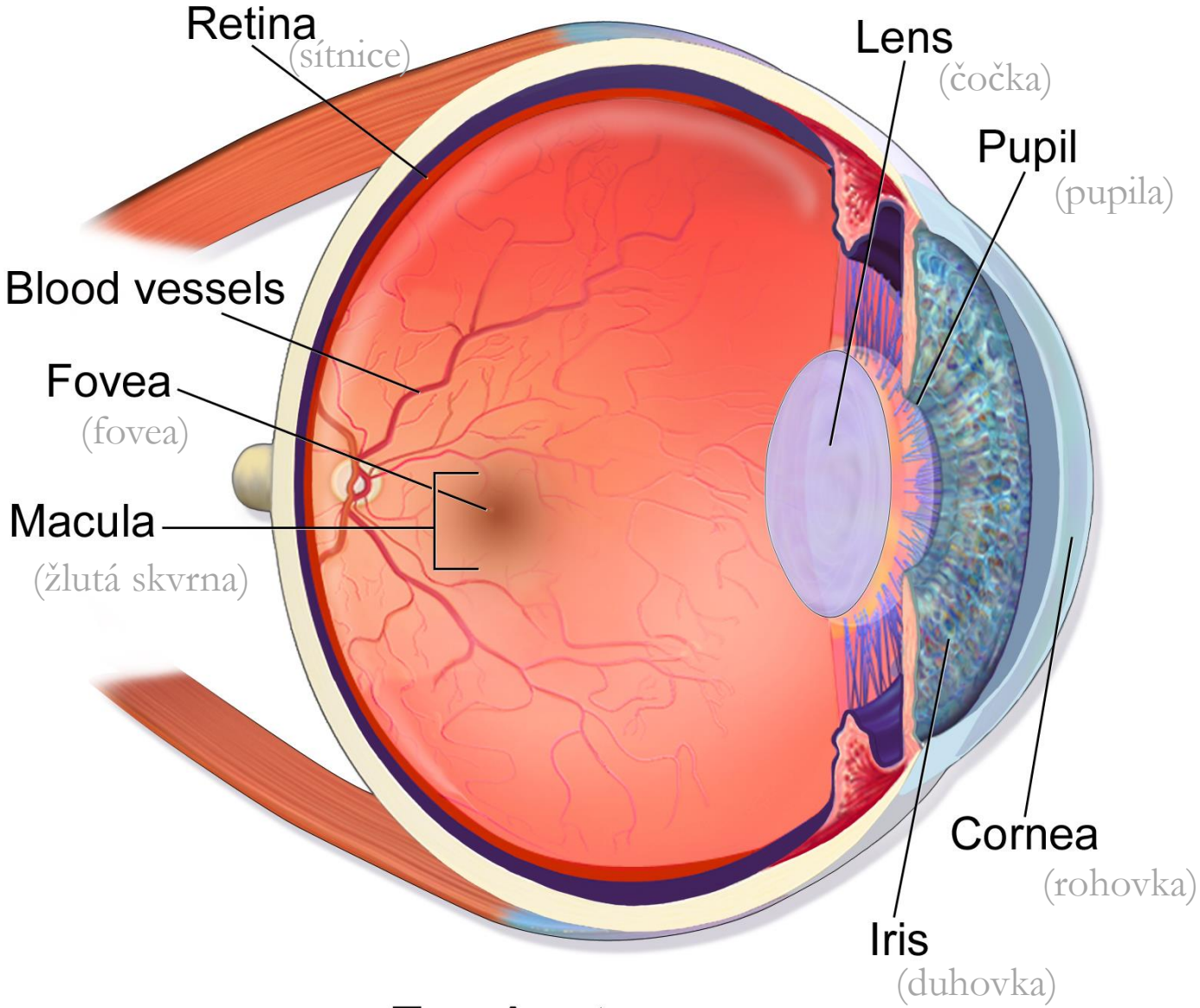
Perception
(cognitive process)

Stimulus → Sensory organ → Perceptual organ



Eye

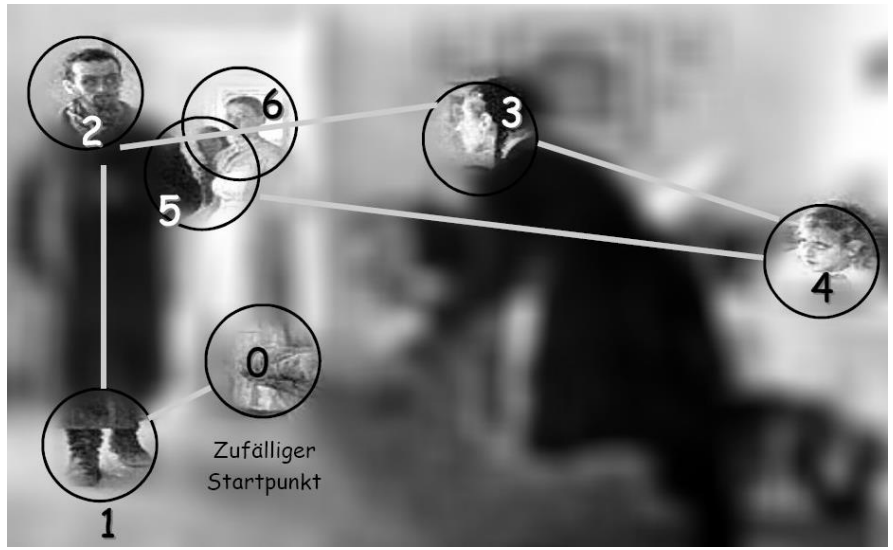
rods (tyčinky), cones (čípky)

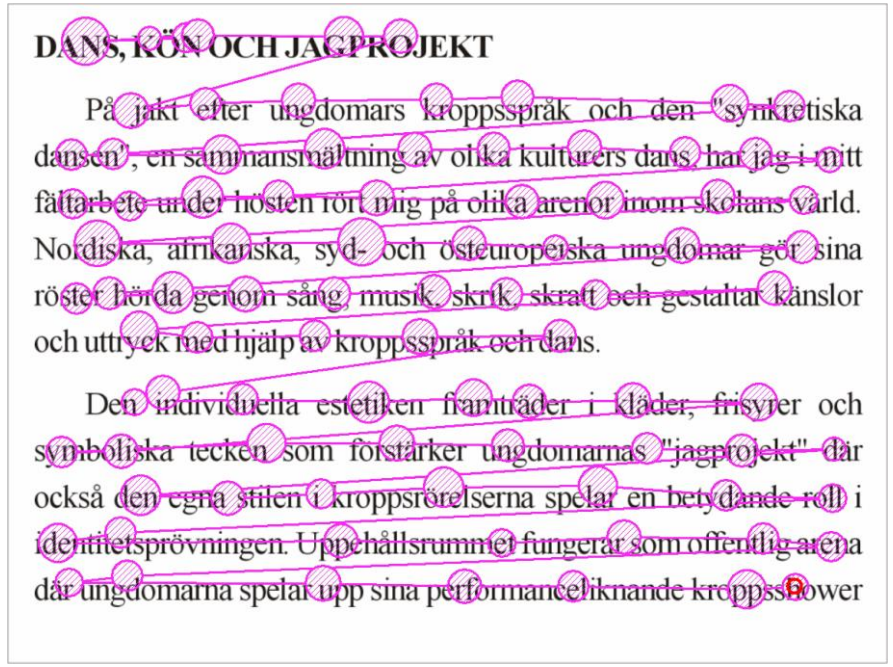


Eye Anatomy

Eye movement

- Since fovea can focus on a limited area at a time, the proces of seeing is not smooth → the sudden change is called **saccade** (saccadic eye movement)
  - Fixation on a particular spot - up to 0.5 s (average 0.25 s)
  - Meanwhile non-foveal parts of the retina survey the full field of vision to chose next spot
  - Jumping to the next point of fixation – 20-40 ms





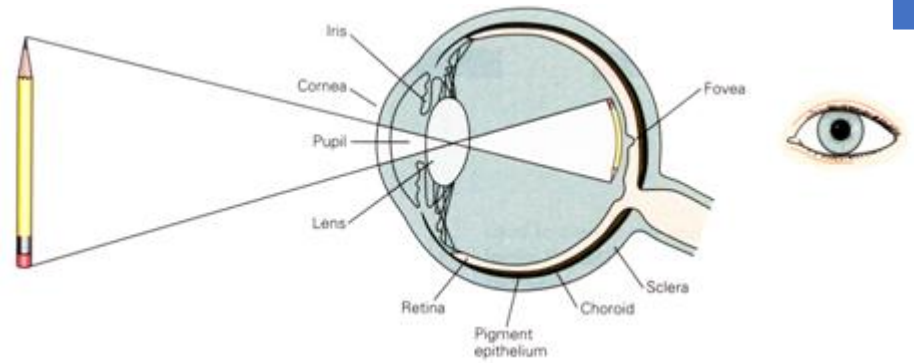
Brain

Sensation
(physical process)

Perception
(cognitive process)

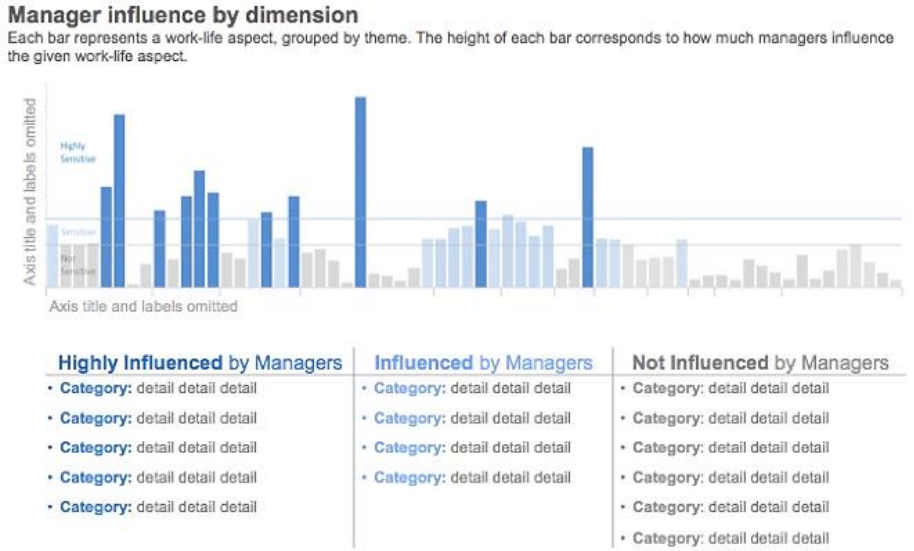
Stimulus → Sensory organ → Perceptual organ

Iconic memory (obrazová paměť) → Working memory (krátkodobá paměť) → Long-term memory (dlouhodobá paměť)



Iconic memory

- Information remains in the iconic memory for less than a second
- Processing in iconic memory is unconscious → ***preattentive processing***
  - Process of recognition, detects attributes such as **color, size, orientation, location** → if something is supposed to stay out it should be encoded using preattentative attributes that contrast with the surrounding information, e.g. red text in the midst of black text or grouping objects together using preattentative attribute (position, color)



source: http://www.storytellingwithdata.com/2011/10/google-example-preattentive-attributes.html

Working memory

- Processing in working memory is conscious → ***attentive processing***
- Data passed from iconic to working memory where they are combined and stored as **visual chunks**
- **Characteristics**
  - **Limited storage capacity**
    - 3-4 chunks
    - A reader can therefore hold only a few chunks of information in her head → a legend of a graph with 10 shapes or colors forces the reader to constantly refer back to the legend
  - **Temporary**
    - If not rehearsed, the chunks stays in the working memory for only a few seconds

Long-term memory

- When it is decided (consciously or unconsciously) that a chunk needs to be stored it is moved (by rehearsal) to long-term memory

- Long-term memory has the ability to recognize images and detect meaningful patterns

Attributes of preattentive processing

- **Visual attributes/encodings** to be perceived by the preattentive processing

Category	Form	Color	Spatial position	Motion
Attribute	Length Width Orientation Shape Size Enclosure …	Hue Intensity	2D/3D position	Direction

Attentive processing – find the number of nines in the list as fast as you can

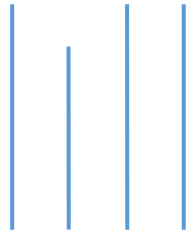
354787654687987184654654654654478913242873
575148642448435545474111123543187584321654
684321354684651313546843513251684651321659
435135143543541321754351351354351213135132
121687465213543517121223313512104768792121
216512135464618435728645646549871731 6

Preattentive processing – find the number of nines in the list as fast as you can

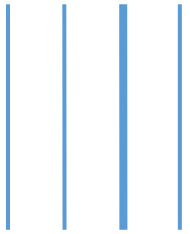
354787654687**9**87184654654654654478**9**13242873
575148642448435545474111123543187584321654
684321354684651313546843513251684651321659
435135143543541321754351351354351213135132
121687465213543517121223313512104768792121
216512135464618435728645646549871731 6

Preattentive attributes of form

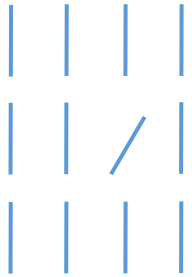
Length



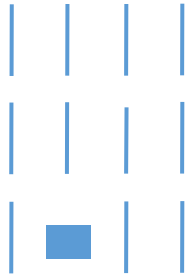
Width



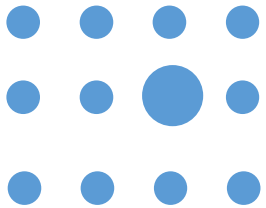
Orientation



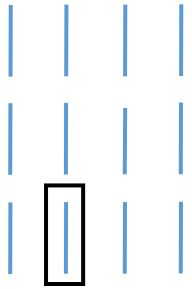
Shape



Size

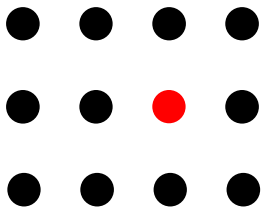


Enclosure

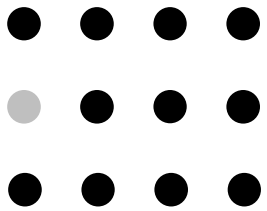


Preattentive attributes of color (1)

Hue



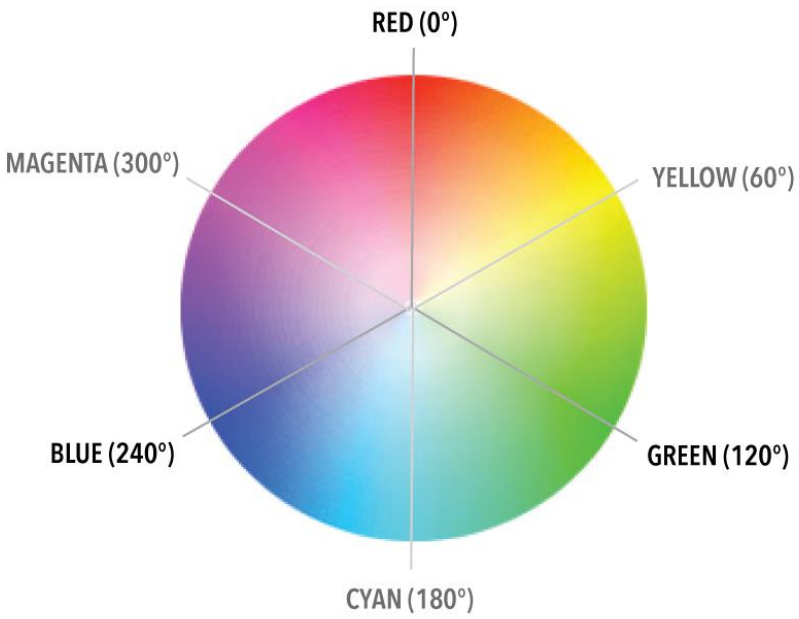
Intensity



Preattentive attributes of color (2)

- Color is made up from three attributes

Hue



Saturation

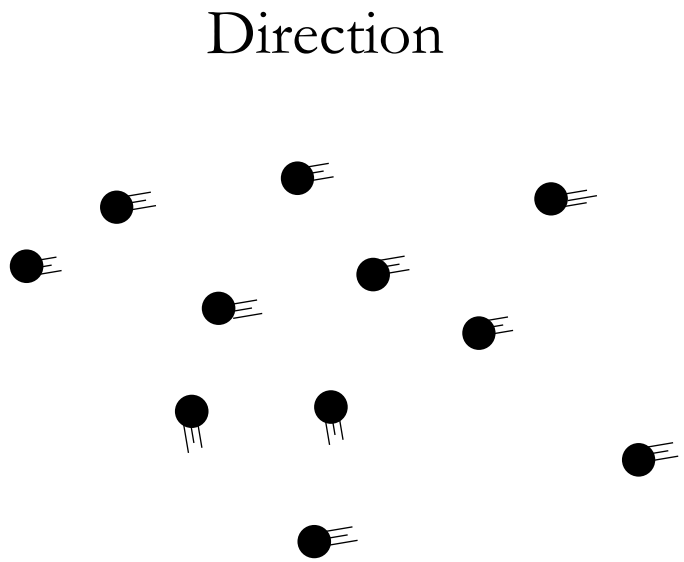
Lightness/Brightness

Intensity applies to both saturation and lightness

Preattentive attributes of spatial position

2D position

Preattentive attributes of motion (1)

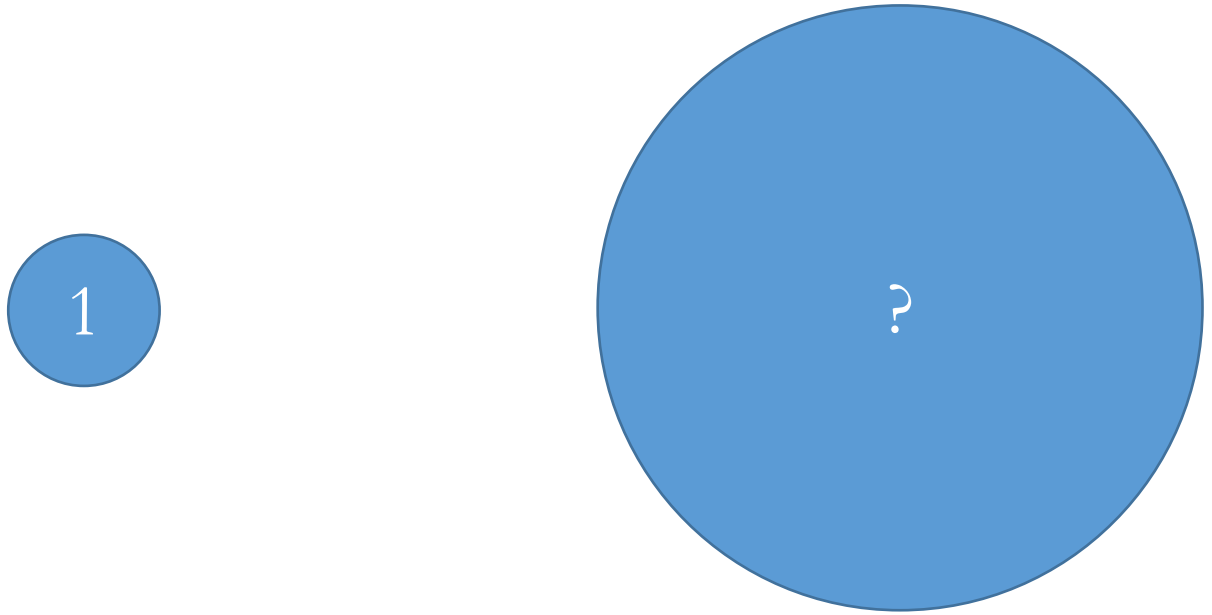


Preattentive attributes of motion (2)

Encoding quantitative values

- **Quantitative** vs **categorical** difference
  - Values represented as **lines of different lengths** are perceived as **quantitatively different** (longer lines greater values)
  - Values represented as **different colors** are only **categorically different** (e.g., red is not "greater" than blue)
    - However, e.g. intensity is perceived quantitatively

Type	Attribute	Quantitatively perceived
Form	Length	Yes
	Width	Yes (limited)
	Orientation	No
	Size	Yes (limited)
	Shape	No
	Enclosure	No
Color	Hue	No
	Intensity	Yes (limited)
Position	2D position	Yes


1
?

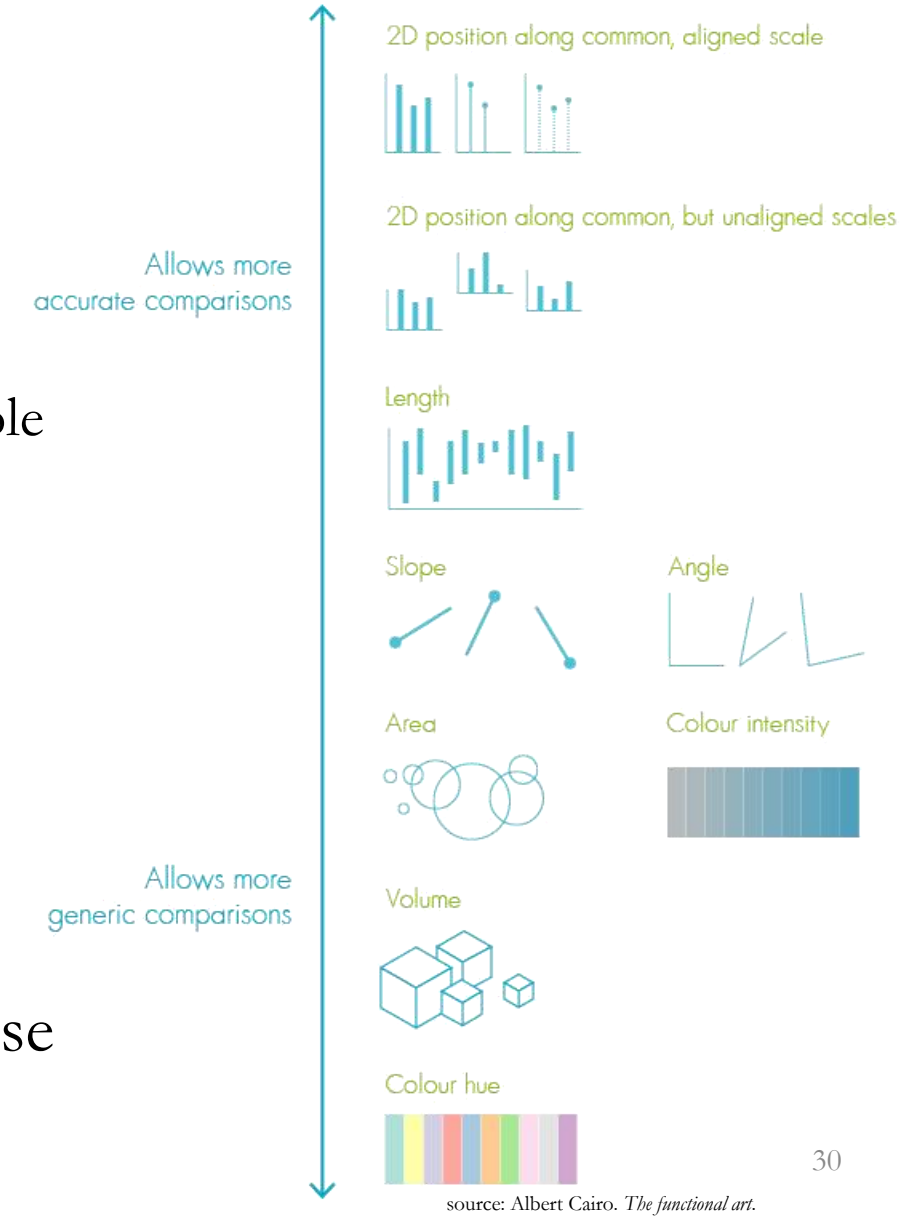
Rankings of visual attributes

- How well people decode visual clues?

1. Position along a common scale (scatter plot)
2. Position on identical but nonaligned scales (multiple scatter plots)
3. Length (bar chart)
4. Angle & slope (pie chart)
5. Area (bubble chart)
6. Volume, density, and color saturation (heatmap)
7. Color hue (newsmap)

William S. Cleveland and Robert McGill. *Graphical Perception and Graphical Methods for Analyzing Scientific Data* Science, 1985, Vol. 229, No. 4716, 828-833

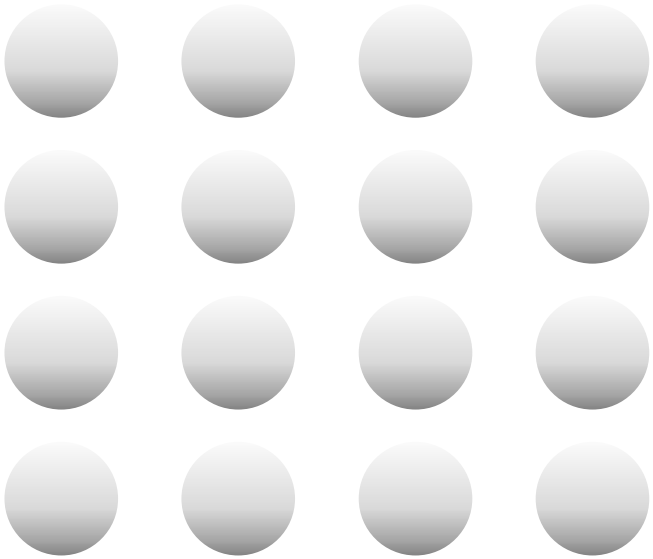
- Therefore, the most important variables are those on the X and Y axis (position)

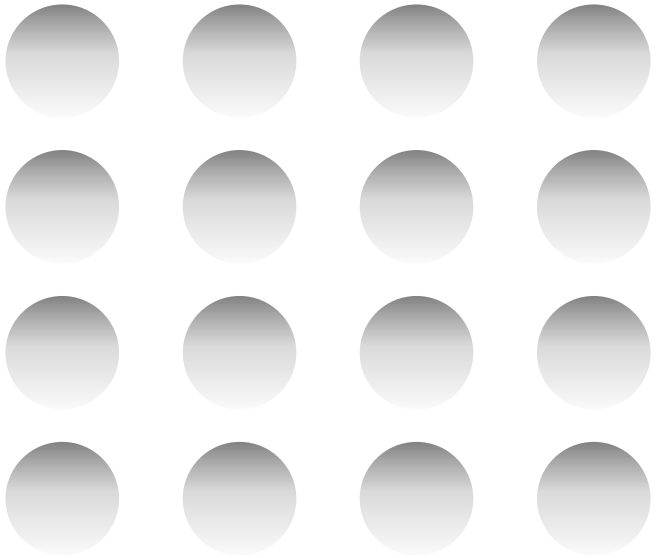


source: Albert Cairo. *The functional art.*

Evolutionary basis of visual perception

- Visual perception is deeply evolutionary ingrained





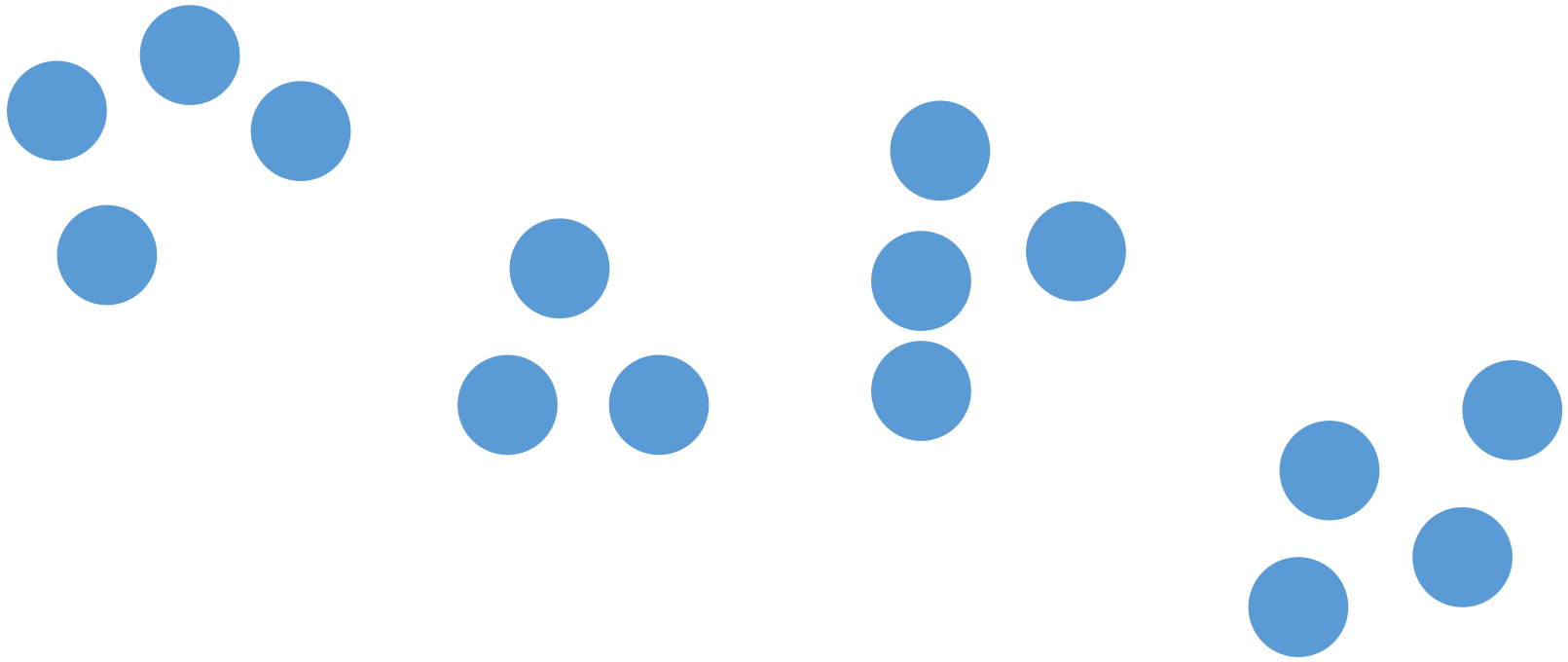
Gestalt principles of visual perception

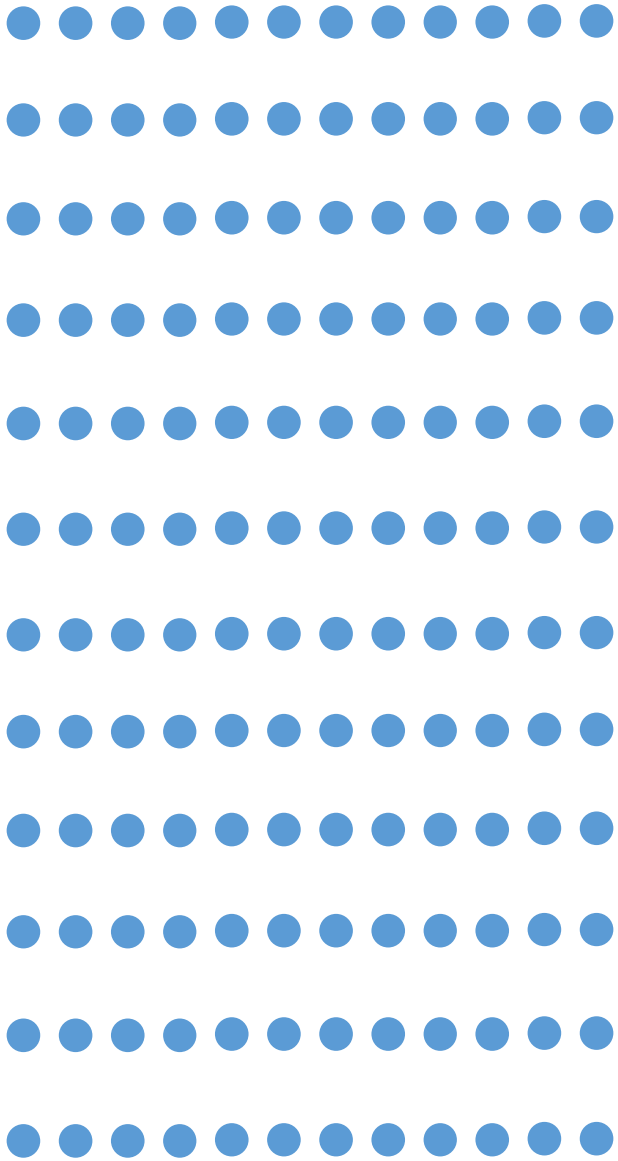
- Gestalt (pattern, shape, form) school of psychology (introduced around 1900)
  - Focus on understanding how we perceive, understand and organize what we see
  - Mind has self-organizing tendencies → **Gestalt laws/principles of grouping**

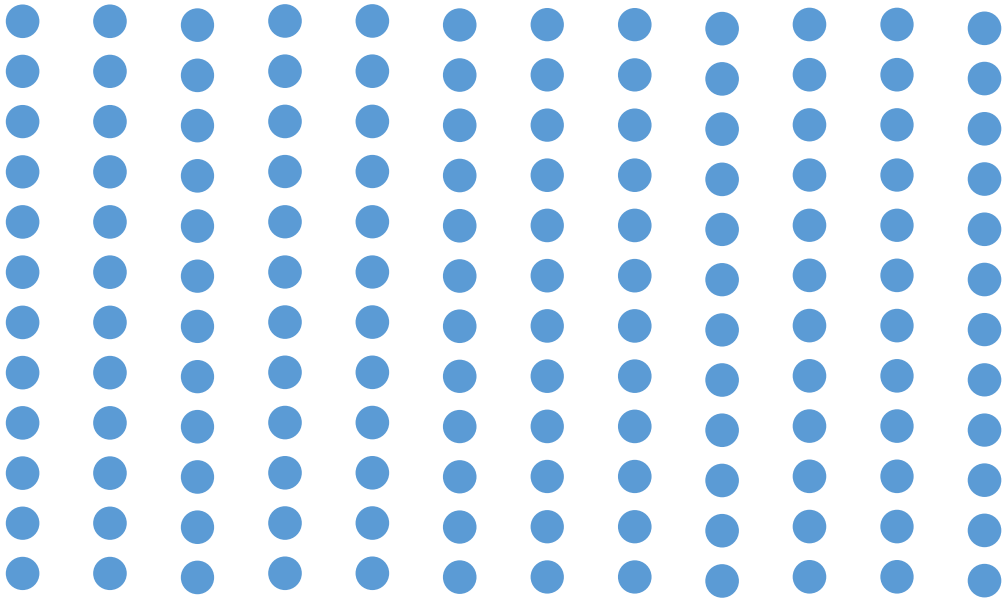
- Principle of proximity
- Principle of similarity
- Principle of enclosure
- Principle of closure
- Principle of continuity
- Principle of connection

Principle of proximity

- **Objects close** to each other are perceived as **forming a group**



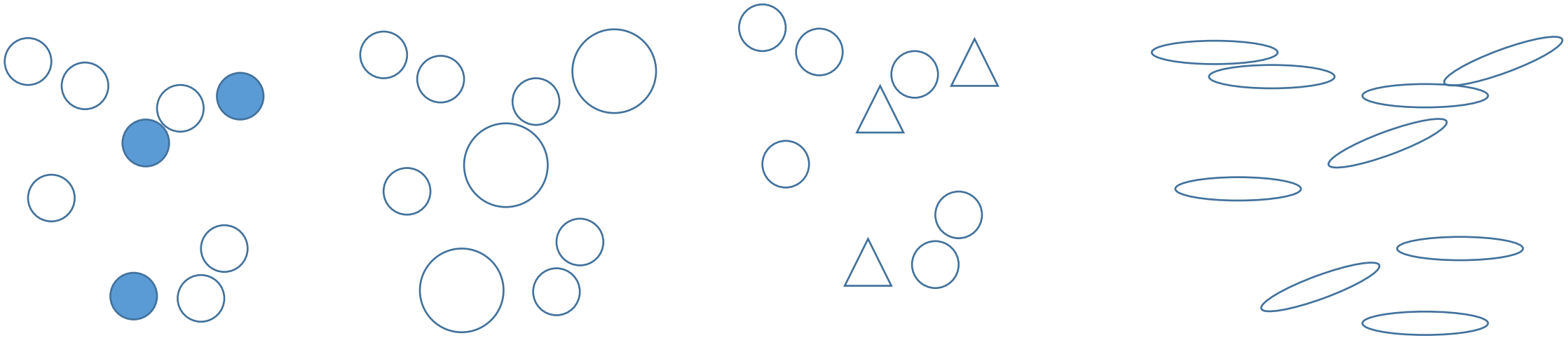




The principle of proximity can be used to direct the reader to scan tables predominantly row or column wise.

Principle of similarity

- Tendency to **group** together object which are **similar** in **color**, **shape** or **orientation**



xxxx xxxx xxxx xxxx xxxx xxxx xxxx xxxx

xxxx xxxx xxxx xxxx xxxx xxxx xxxx xxxx

xxxx xxxx xxxx xxxx xxxx xxxx xxxx xxxx

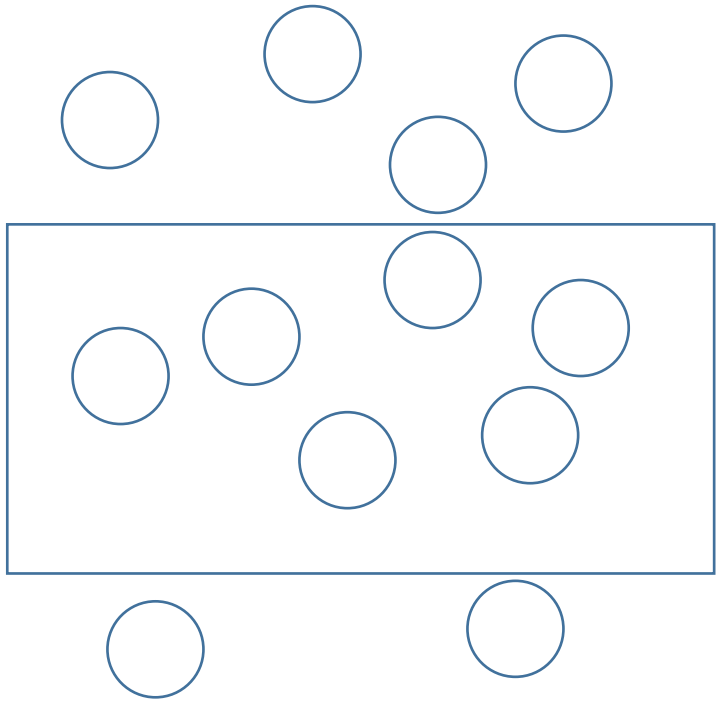
xxxx xxxx xxxx xxxx xxxx xxxx xxxx xxxx

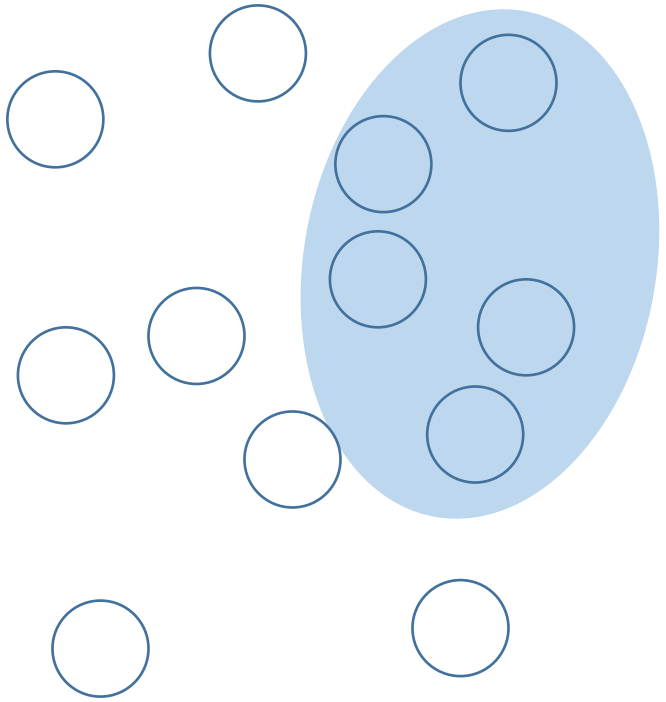
xxxx xxxx xxxx xxxx xxxx xxxx xxxx xxxx

xxxx xxxx xxxx xxxx xxxx xxxx xxxx xxxx

Principle of enclosure

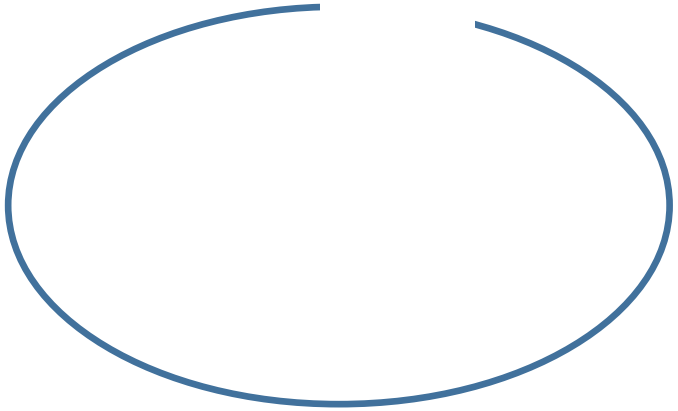
- We perceive objects belonging **together** when they are somehow **enclosed**





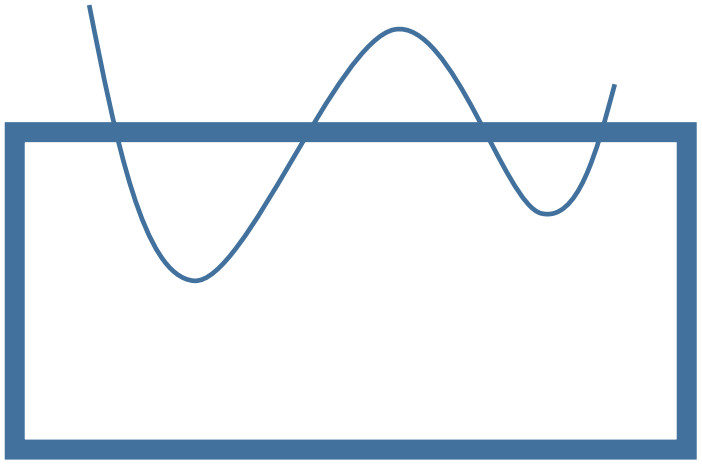
Principle of closure

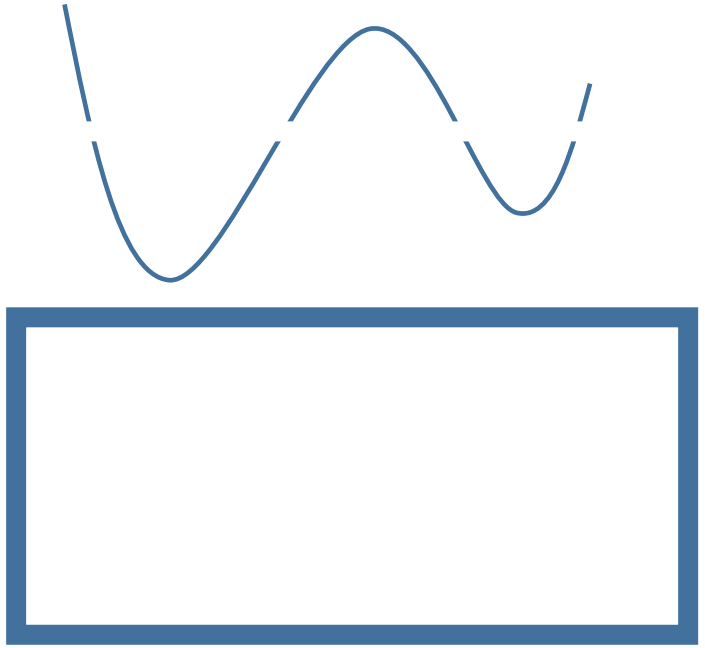
- If there is an ambiguous stimuli we will try to **eliminate the ambiguity**
- We prefer to see objects as **closed**, **complete** and **regular**

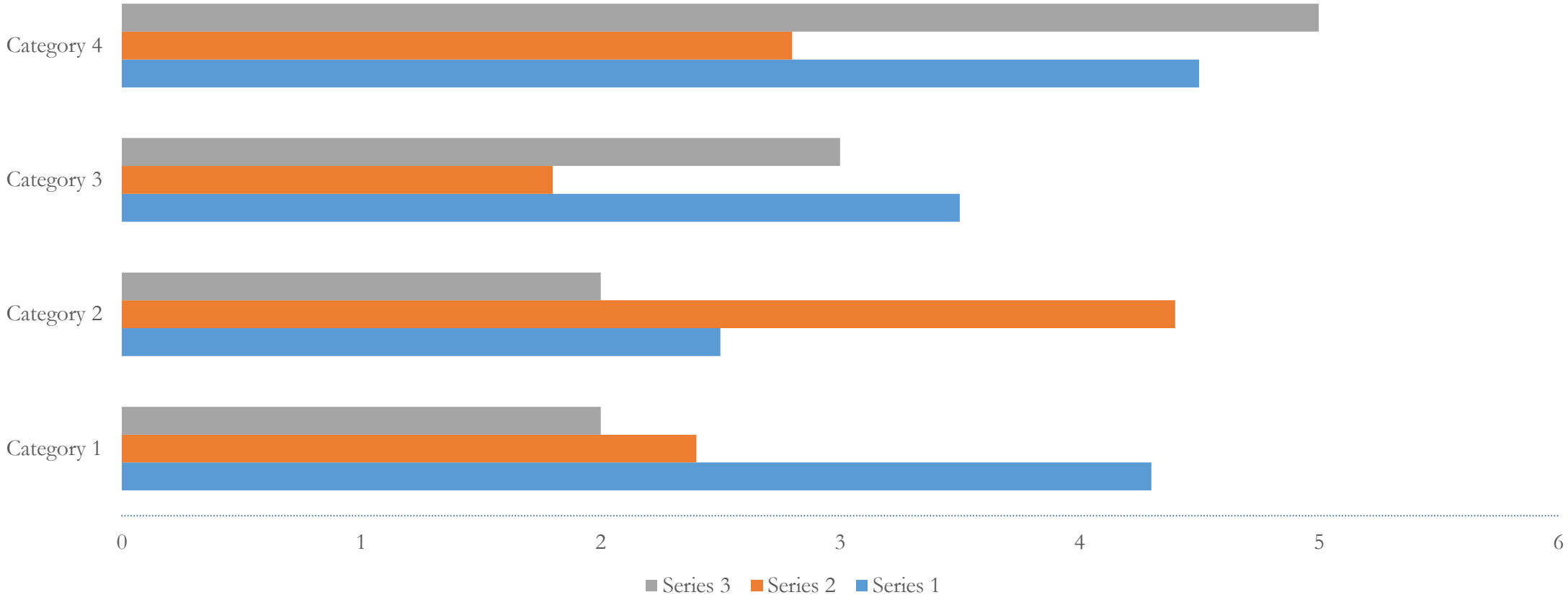


Principle of continuity

- We perceive objects as belonging together, forming a **whole**, if they are **aligned or connected** to one another



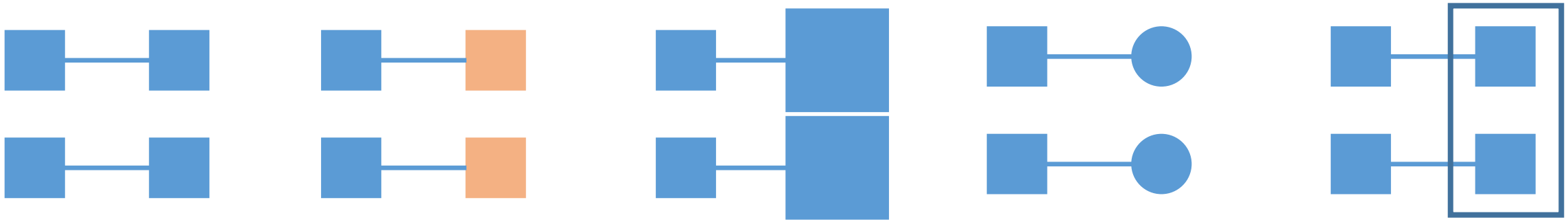


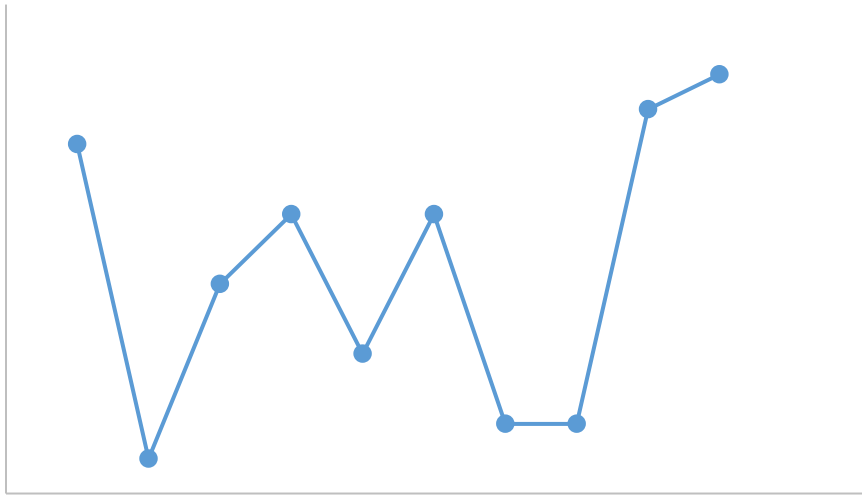


Due to the principle of continuity there is no need for the "0" line

Principle of connection

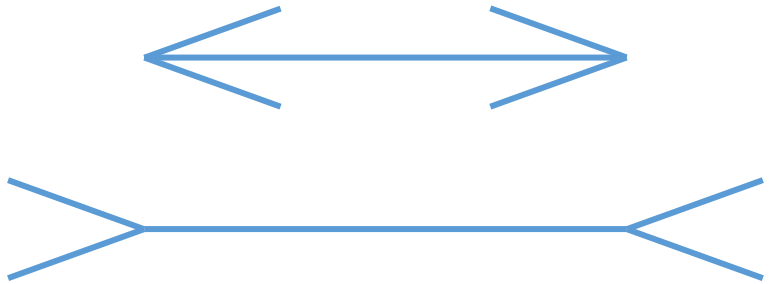
- **Connected** objects are perceived as **part of a group**
- Connection exercises greater power than proximity or similarity but less than enclosure





Effects of context

- Our visual **senses** are designed to perceive **differences** in values **rather than absolute values**

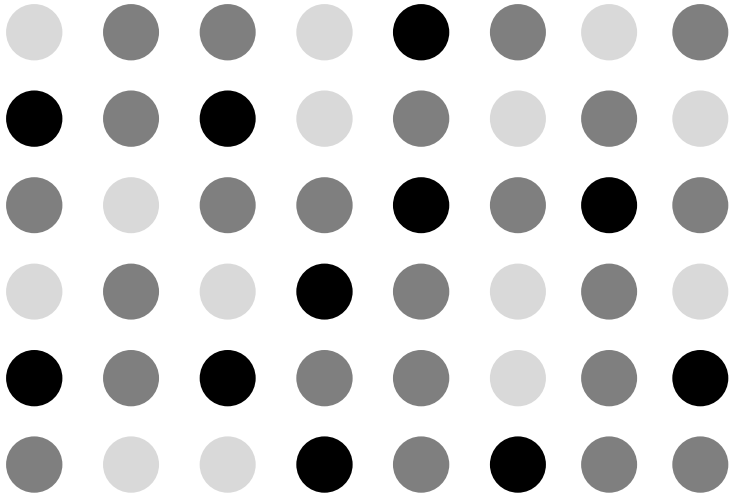


Limits to distinct perception (1)

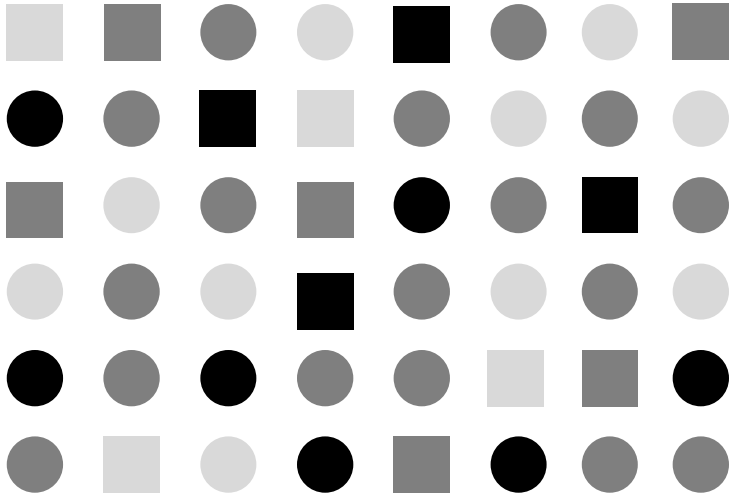
- **Too much** visual attributes or values per attribute **can harm**
  - "*It is simple to spot a single hawk in a sky full of pigeons, but it would be more difficult if the sky contained more types of birds*" (Ware, 2004)
  - Using larger number of values **forces readers** to use the slower **attentive processing** which allows to store only up to four distinctive values at a time

Limits to distinct perception (2)

- **Preattentive processing** usually **cannot** handle **more than one** visual **attribute** of an object at a time



Focus on black objects

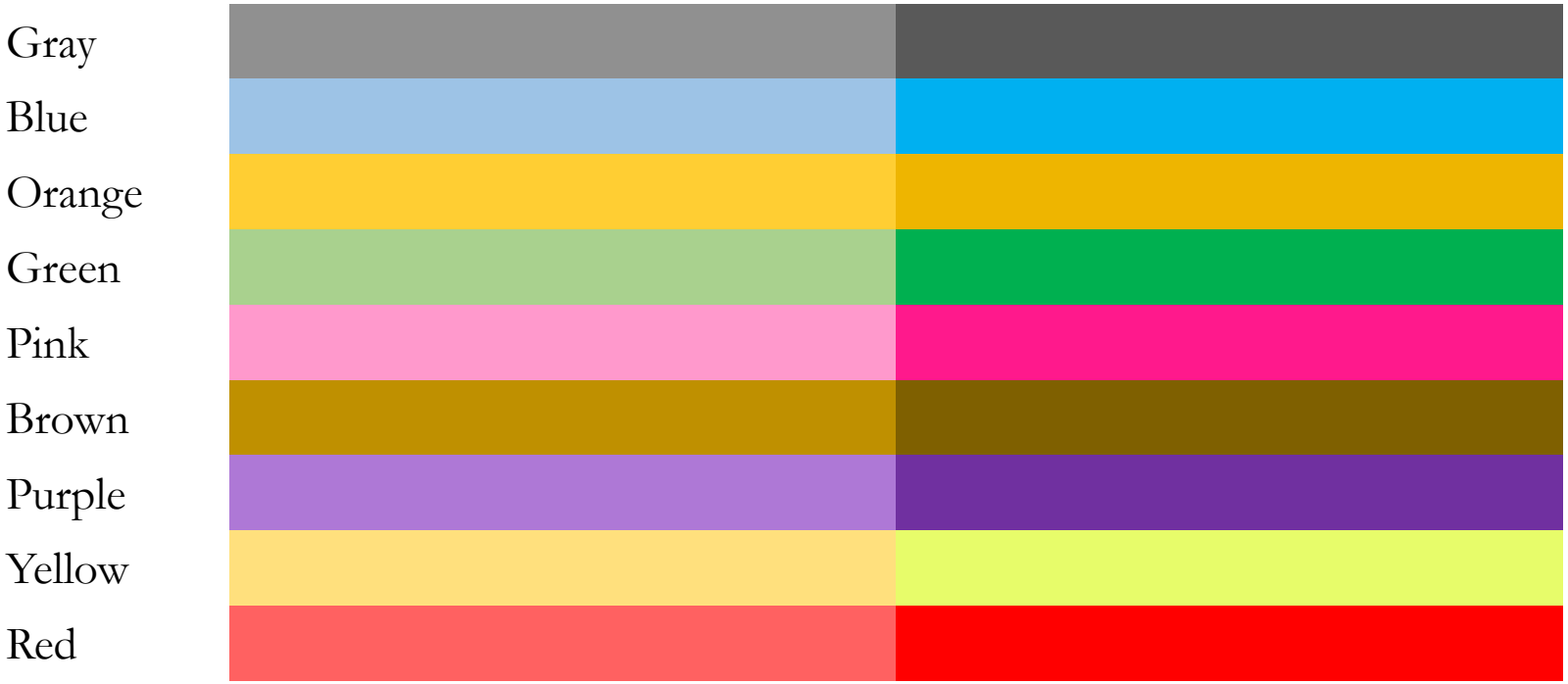


Focus on white objects

Focus on white squares

Limits to distinct perception (3)

- There are **nine hues** that are easy to recognize

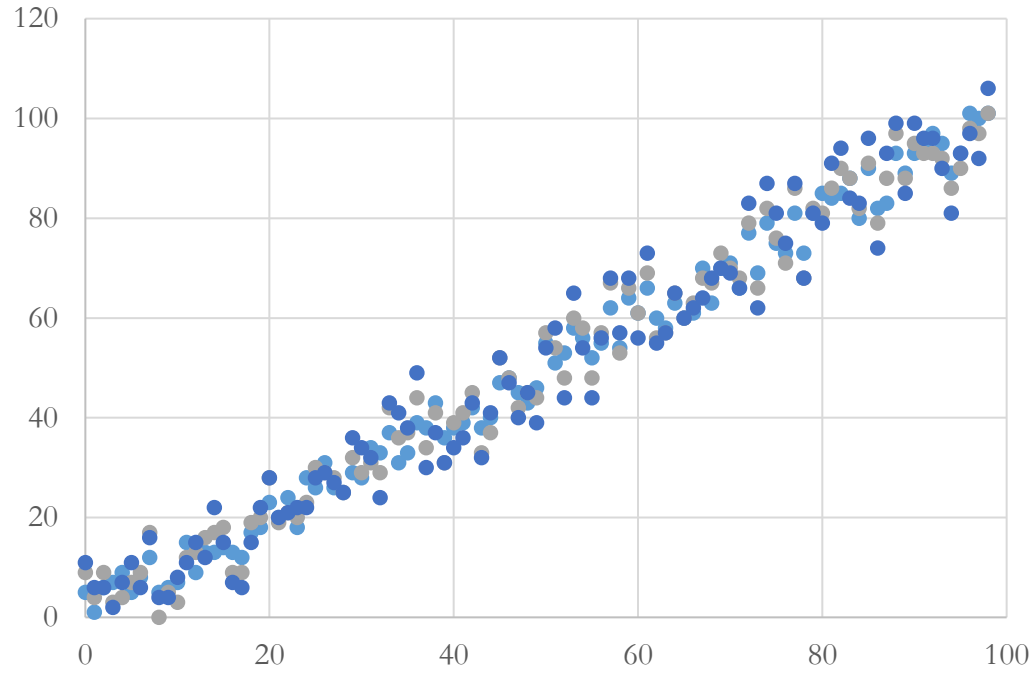


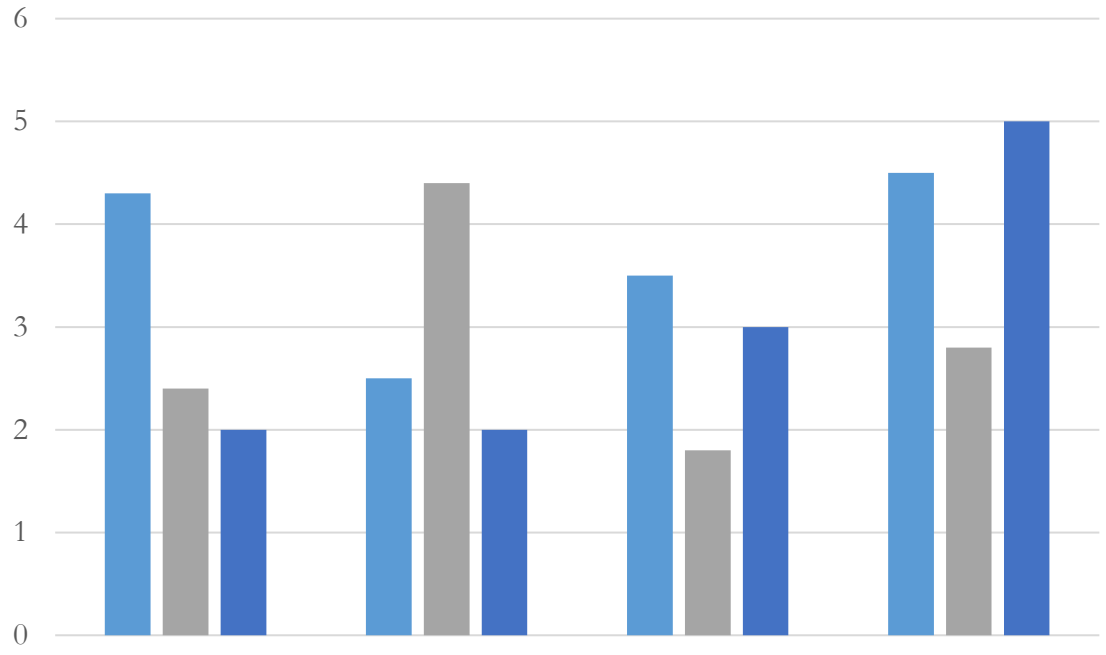
Soothing colors, suitable for tables and graphs.

Vibrant colors, suitable for highlighting.

Limits to distinct perception (4)

- The **ability to distinguish colors decreases** along with the **sizes** of objects → small objects (points in graph) should be darker than large objects such as bars





Data types

Quantitative data

- Deals with **numbers**
- Can be **measured**
- Stored in **numeric variables**
- Length, height, area, volume, weight, speed, time, temperature, humidity

Categorical/qualitative data

- Deals with **descriptions**
- Can be observed but **not measured**
- Stored in **categorical variables**
- Gender, color, texture, taste, appearance

Quantitative

- 120 x 100 cm
- Weights 0,5kg
- 2 people
- 1 animal

Qualitative/categorical

- Aquarelle
- Darker colors
- Contains text
- Masterful brush strokes

Quantitative relationships

- **Quantitative stories** are about **relationships** which, in turn, determine the type of visualization to relay the story (table, graph, diagram, …)

Quantitative data vs categorical data

Quantitative information	Relationship
Units of product sold per geographical location	Sales related to geography
Revenue by quarter	Revenue related to time
Expenses by department and month	Expenses related to organization structure and time
A company's market share compared to that of its competitors	Market share related to companies
The number of employees who received each of the five possible performance ratings during the last annual performance review	Employee counts related to performance ratings

Relationships between categorical items

Categorical items used to label corresponding measures relate to one another in the following ways

Nominal *(jmenné)*

- Values in a single category are discrete and have no intrinsic order
- Sales in regions (East, West, North, South)

Ordinal *(pořadové)*

- The categorical items have order
- Size (small, medium, large)

Interval *(měřitelné, intervalové)*

- Categorical items consist of a sequential series of numerical ranges
- Order size ($0-$1,000;$1,000-$2,000;>$2,000)

Hierarchical *(hierarchické)*

- Involves multiple categories in the parent-child relation (tree structure)
- Organizational structure (division → department → group)

Relationships between quantities

- Categorical items can relate by quantitative values associated with them

Ranking	Ratio	Correlation
• Order of the categorical items based on the associated quantitative values • Sales orders (top five orders for the current quarter based on revenue)	• Compares two quantitative values by dividing one by the other • Value of a single categorical item compared to the sum of the entire category (market share) • Measure of change (expenses from one month to the next)	• Compares two paired sets of quantitative values • Relation between values (number of years on a job and productivity)

Numbers that summarize

- Ways to summarize/aggregate data → **descriptive statistics**
- Reduces large sets of data allowing to comprehend the story

- Measures of **average** (central tendency, center)
- Measures of **variation**
- Measures of **correlation**
- Measures of **ratio**

Measures of average (1)

(Arithmetic) Mean

- **Sum** of all the values **divided** by the **number** of values
- Measure of center taking into account all values → prone to be **influenced** by **extreme values**

- E.g., in case of salaries it can be used to show comparative impact of departments of a company on expenses

Median

- Value from the **middle** of the (sorted) set
- Expresses the **typical values**

- E.g., in case of salaries it can be used to show typical salary (per department)

Measures of average (2)

Mode

- The specific **value** that appears **most often** in the set
- If there are more values like this, the set is multimodal
- If there is no such value, the set does not have a mode

Midrange

- The value **midway** between **highest** and **lowest** value
- Quick estimate of center
- Very sensitive to extremes (if the distribution is not uniform)

Measures of variation (1)

- Presents the degree into which values vary

Spread	Standard deviation
• **Difference** between the **lowest** and the **highest** value • Relies on too little information • Affected by extreme values	• Measures **variation** in a set **relative to mean** • The higher the number of values the less it is prone to bias due to the extreme values

Measures of variation (2)

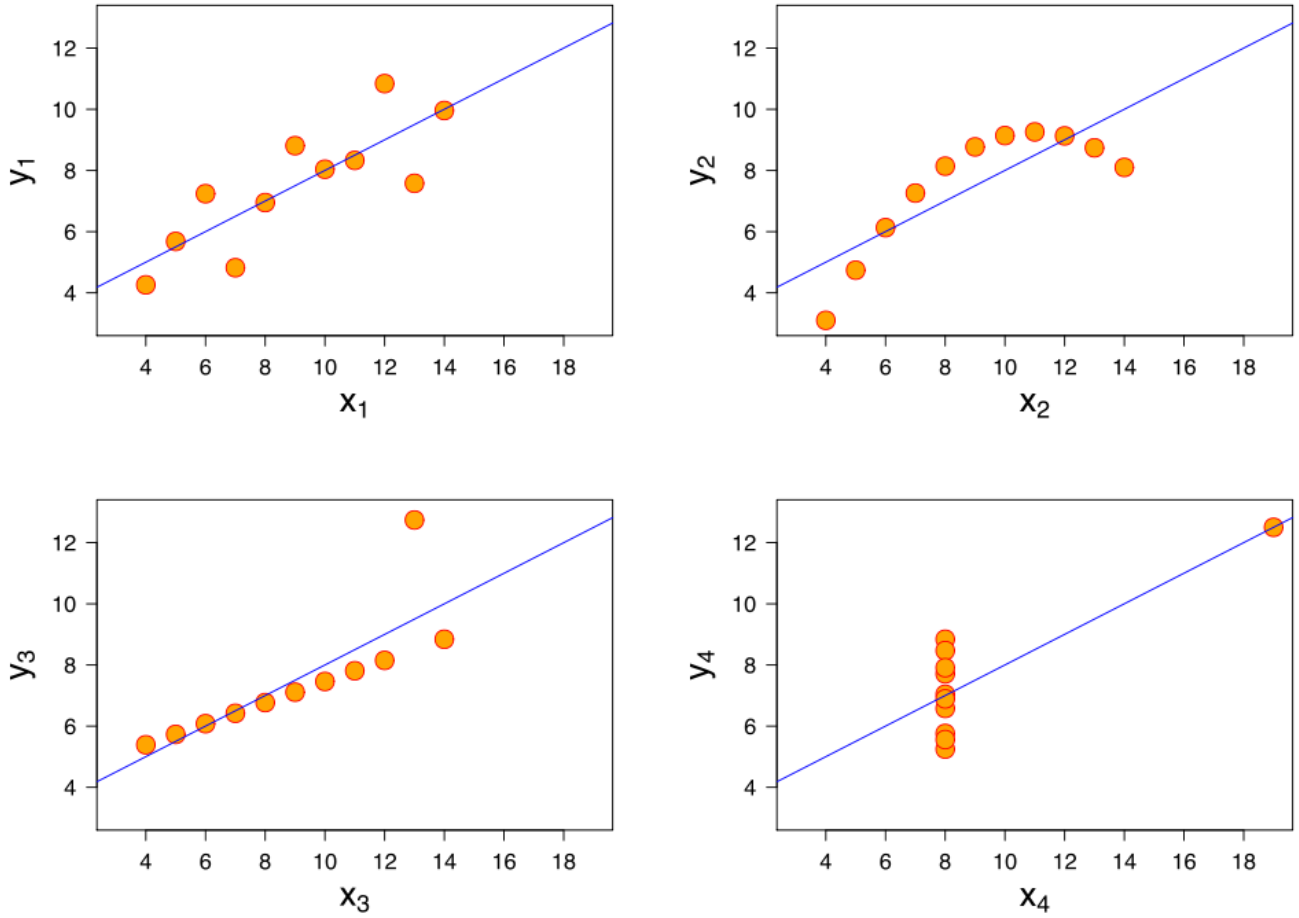
- High variation in time it takes to manufacture products, answer phone calls, or resolve technical calls
  - Does it indicate problems in training, process design, or systems?

- Variation in departmental expenses.
  - Do some departments manage to keep their expenses much lower than others? Why?

- Variation in food quality of a restaurant reported by customers.
  - Does the variation relate to the cook?

Measures of correlation

- The simplest relation we measure is linear correlation being commonly expressed in terms of **the (Pearson) correlation coefficient**
- Ranges between -1 (strongest negative correlation) and 1 (strongest positive correlation)



corr(x,y) = 0,816

Measure of ratio

- Measures relation between a single pair of values (unlike correlation)
- Can be expressed in the following ways

Sentence

Two out of every five customers who access our website place an order

Fraction

2/5

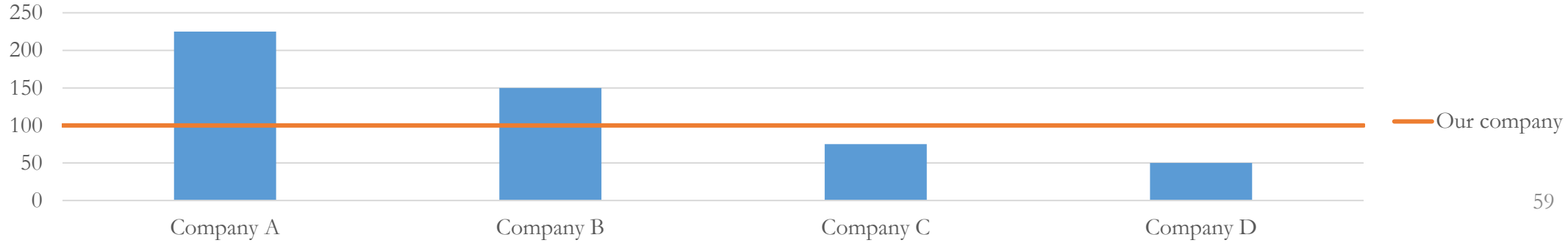
Rate

0.4 (i.e., the result of division 2/5)

Percentage

40% (i.e., the 0.4 rate multiplied by 100)

- Special use case is setting one of the values constant → baseline to which other value are compared (set to 1 or 100%)



General data sources

- World Bank
- EU Open Data Portal
- Eurostat
- data.gov.uk
- U.S. Government's open data
- OECD
- Knoema
- OpenData.cz
- Data Portals
- Tableau
- ManyEyes
- ….

“Big data” repositories

- Click (2.5 TB)
- Tiny images (227 GB)
- Wikipedia edits (2GB)
- 1000 genomes project (260 TB)
- ….

Machine learning related repositories

- Kaggle datasets
- Kdnuggets datasets
- mldata
- http://archive.ics.uci.edu/ml/
- ….

Sources

- Stephen Few (2012) Show Me the Numbers – Designing Graphs and Tables to Enlighten