

What is Visualization

Data Management and Visualization



SoftEng
<http://softeng.polito.it>

Version 2.3.0
© Marco Torchiano, 2021





This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

To view a copy of this license, visit

<http://creativecommons.org/licenses/by-nc-nd/4.0/>.

You are free: to copy, distribute, display, and perform the work

Under the following conditions:



Attribution. You must attribute the work in the manner specified by the author or licensor.



Non-commercial. You may not use this work for commercial purposes.



No Derivative Works. You may not alter, transform, or build upon this work.

- For any reuse or distribution, you must make clear to others the license terms of this work.
- Any of these conditions can be waived if you get permission from the copyright holder.

Your fair use and other rights are in no way affected by the above.

Topics

- Visualization literacy
 - ◆ Visual perception
 - ◆ Graph design
- Visualization skill
 - ◆ Tool:  + a b | e a u®
 - ◆ Practice with different visualization problems and graph types

Exam [5 points + Theory]

- Assessment
 - ◆ Question [0.25]
 - ◆ Data [1.25]
 - ◆ Visual
 - Proportionality [0.75]
 - Utility [0.75]
 - Clarity [0.5]
- Redesign [0.25 + 1.25]

Definition

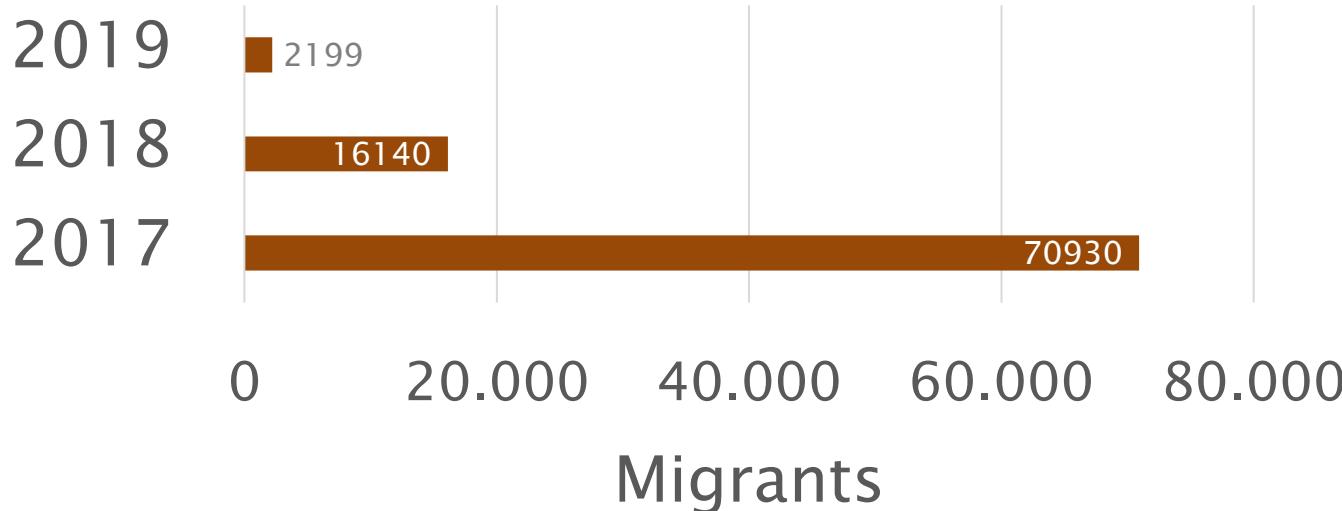
Visualization:

Usage of visual features to encode data in order to convey useful information



WHY VISUALIZATION?

Migrants arrived in period January – June



http://www.interno.gov.it/sites/default/files/cruscotto_statistico_giornaliero_19-06-2019.pdf

The accidents at work happened and reported to Inail in first quarter 2019 have been 131 thousand (109 thousand at work and 22 thousand while traveling), on the rise by 1,7% (+2 thousand reports) with respect to first quarter 2018

https://www.istat.it/it/files/2019/06/NotaTrimestrale–Occupazione-I_2019.pdf

Motivation

Information retrieval

- After 3 days
 - ◆ Text alone: 10%
 - ◆ Text + visuals: 65%

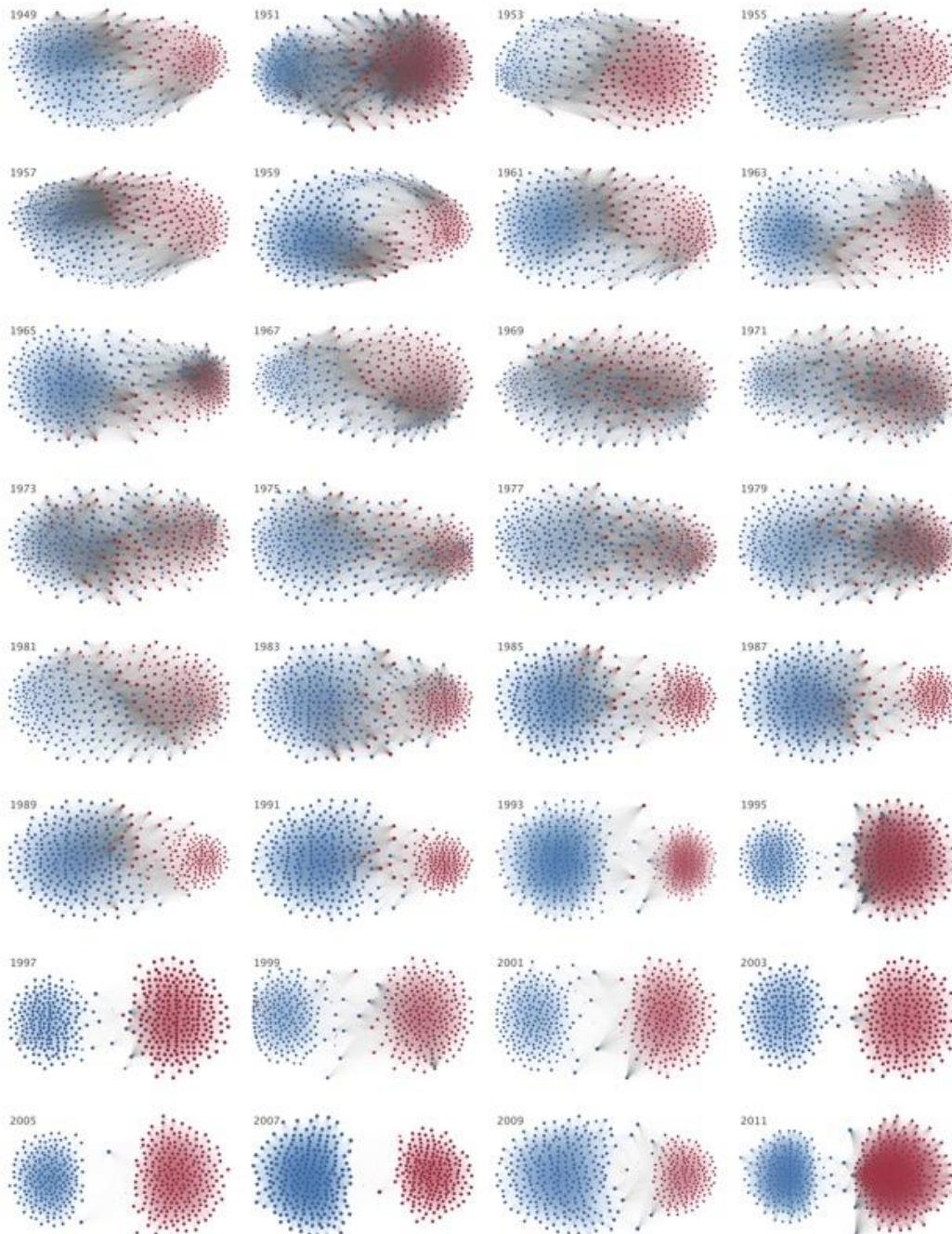
[John Medina, Brain Rules, 2008]

Motivation

Information retrieval

Information density

- In principle every single pixel in an image could encode a datum
 - ◆ Screen (1024x768) ~ 1 M pixels
 - ◆ 1 M characters ~ 250 pages



Motivation

Information retrieval

Information density

Information context

Visualization compares multiple values
and puts the information into context.
A single number means nothing.

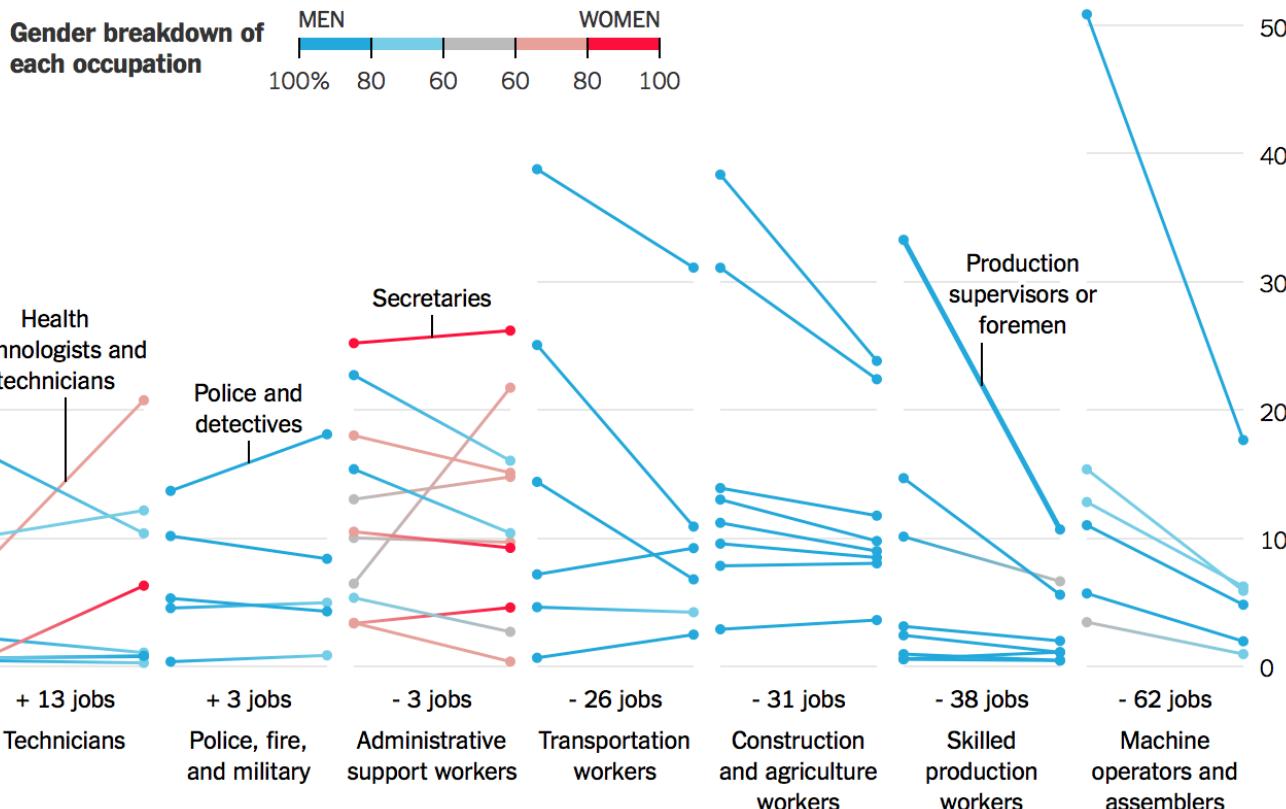
[Randy Krum presentation at Malofiej 23 (March 2015)]

The Changing Nature of Middle-Class Jobs

By GREGOR AISCH and ROBERT GEBELOFF FEB. 22, 2015

The types of jobs that pay middle-class wages — between \$40,000 and \$80,000 in 2014 dollars — have shifted since 1980. Fewer of these positions are in male-dominated production occupations, while a greater share are in workplaces more open to women.

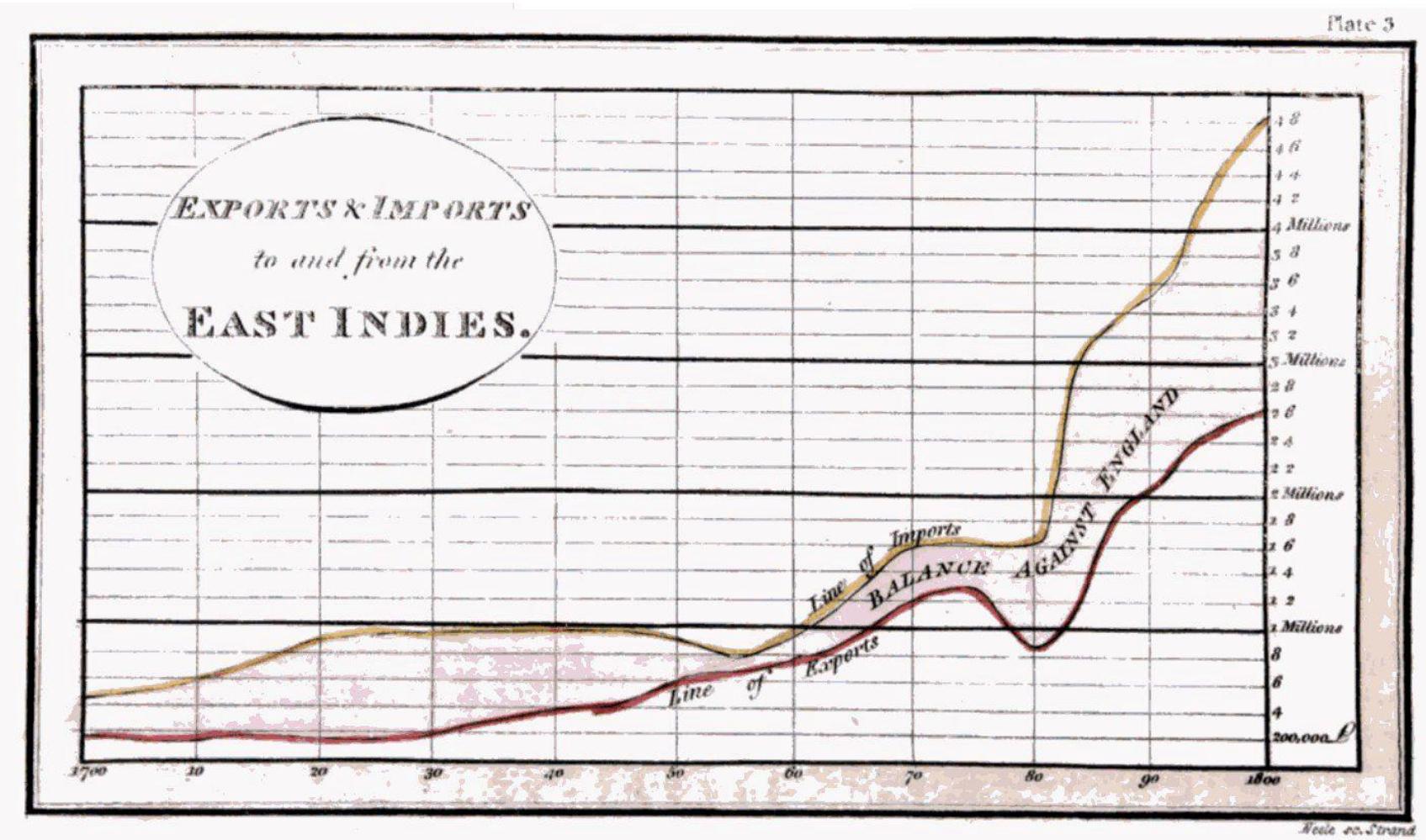
RELATED ARTICLE



HISTORY

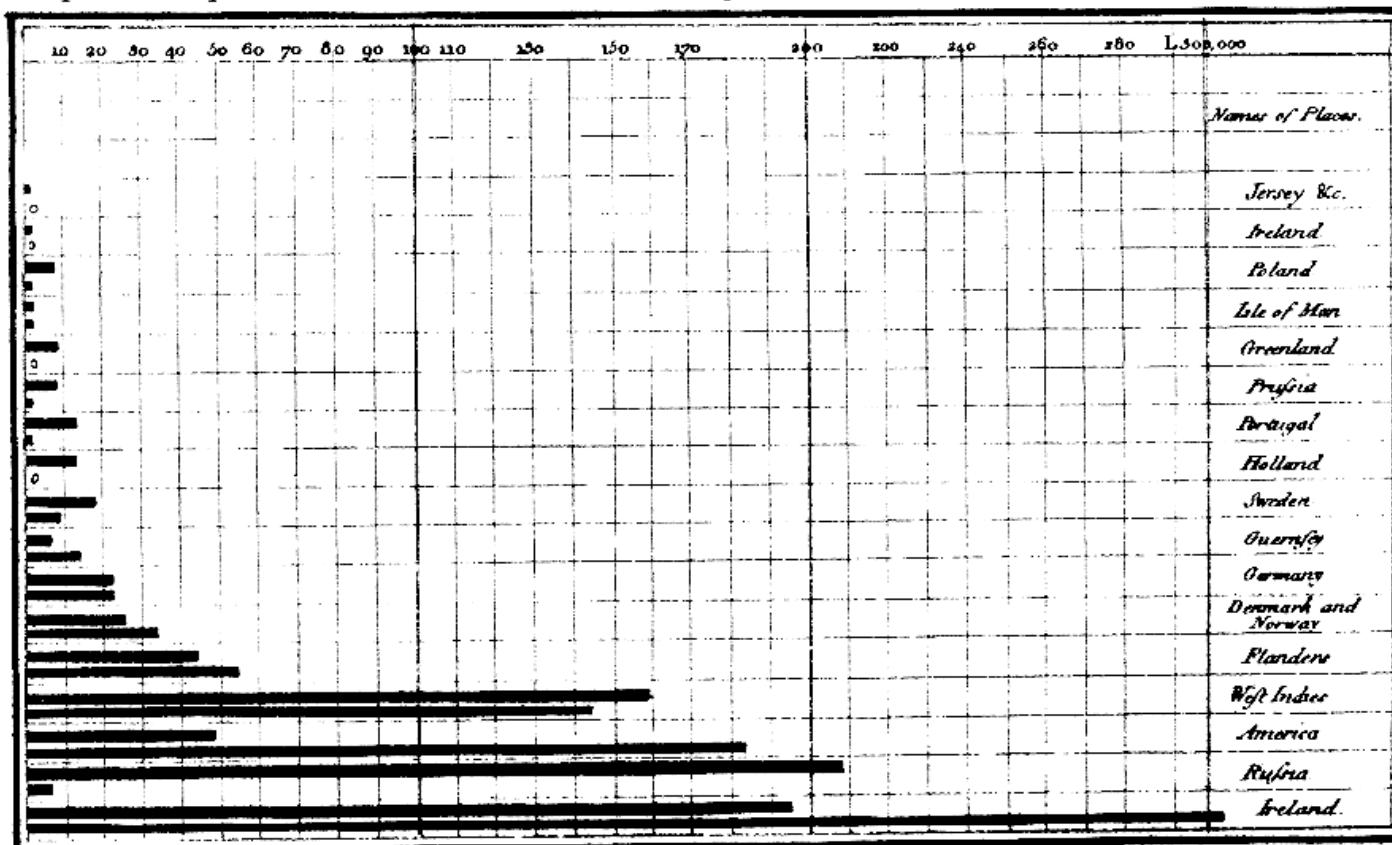
William Playfair

Plate 3



W.Playfair, The Commercial and Political Atlas, London 1786

Exports and Imports of SCOTLAND to and from different parts for one Year from Christmas 1780 to Christmas 1781.



The upright divisions are Ten Thousand Pounds each. The Black Lines are Exports the Ribbed lines Imports.
Published in the Commercial Atlas of 1781 by W. Playfair.

No. 352, Strand, London.

W. Playfair, The Commercial and Political Atlas, London 1786

Charles Joseph Minard

*Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813.
Dessinée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.*

Paris, le 20 Novembre 1869.

Les nombres d'hommes présents sont représentés par les largeurs des zones colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zones. Le rouge désigne les hommes qui entrent en Russie; le noir ceux qui en sortent. Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M. M. Chiers, de Léger, de Férouzat, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre.

Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davout, qui avaient été détachés sur Minsk et Mohilow en se rejoignant vers Orsha et Vitebsk, avaient toujours marché avec l'armée.

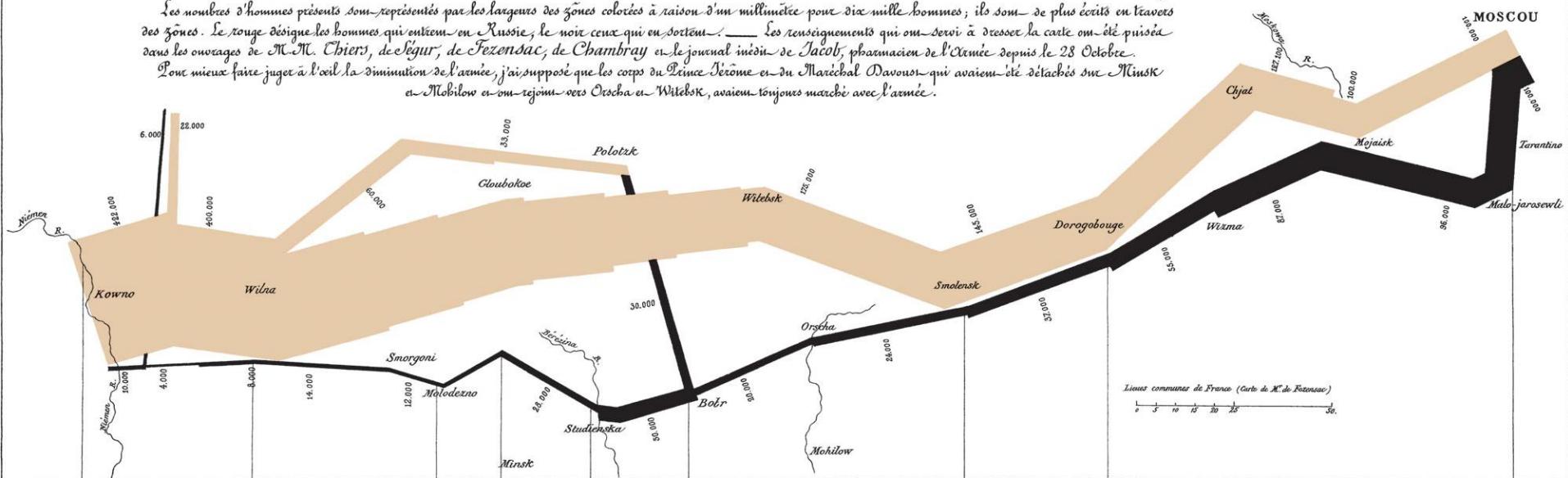
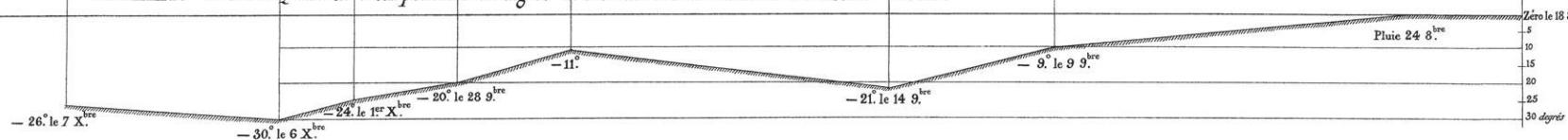


TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.

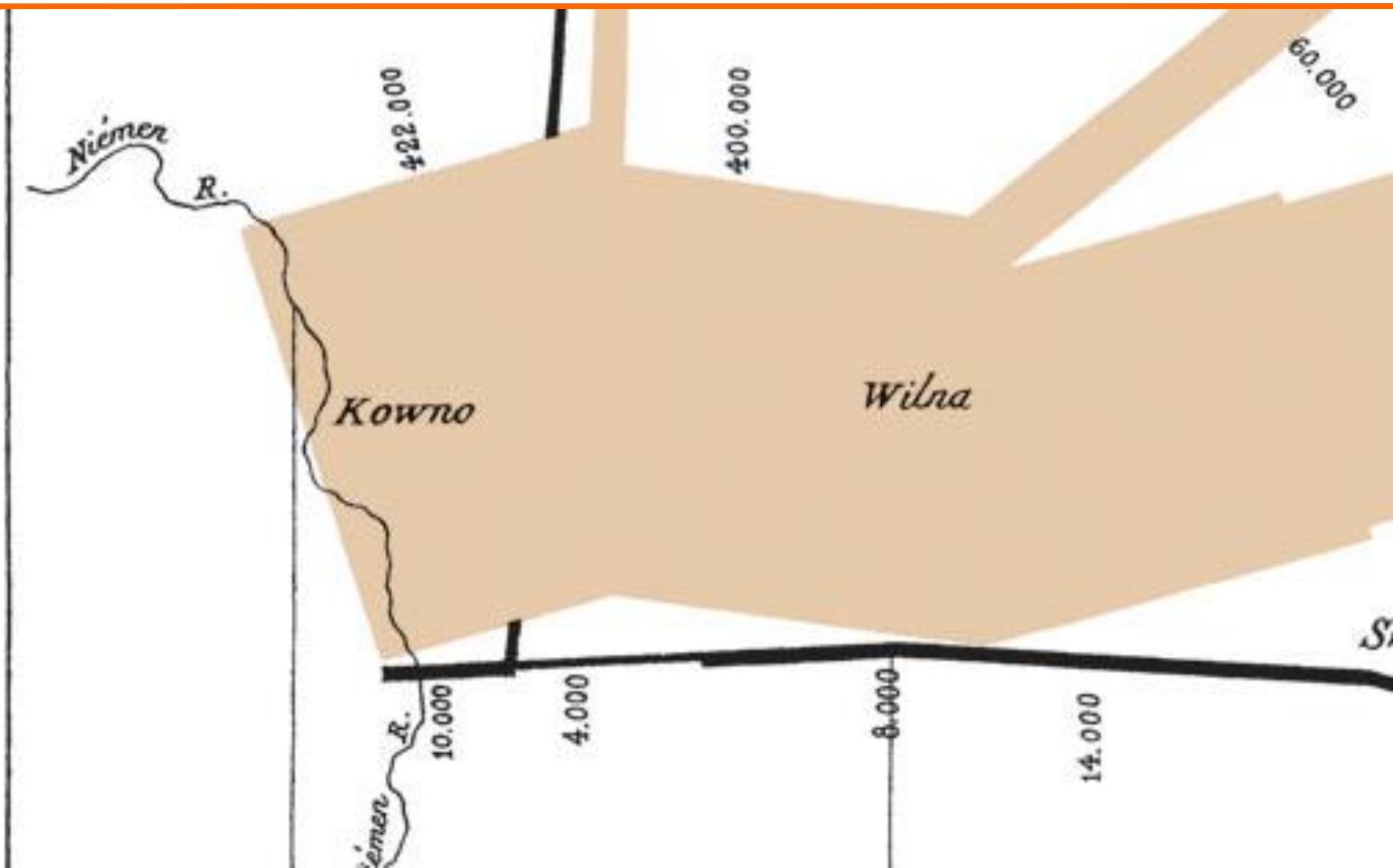


Autog. par Regnier, S. Pas. S^e Marie S^e G^e à Paris.

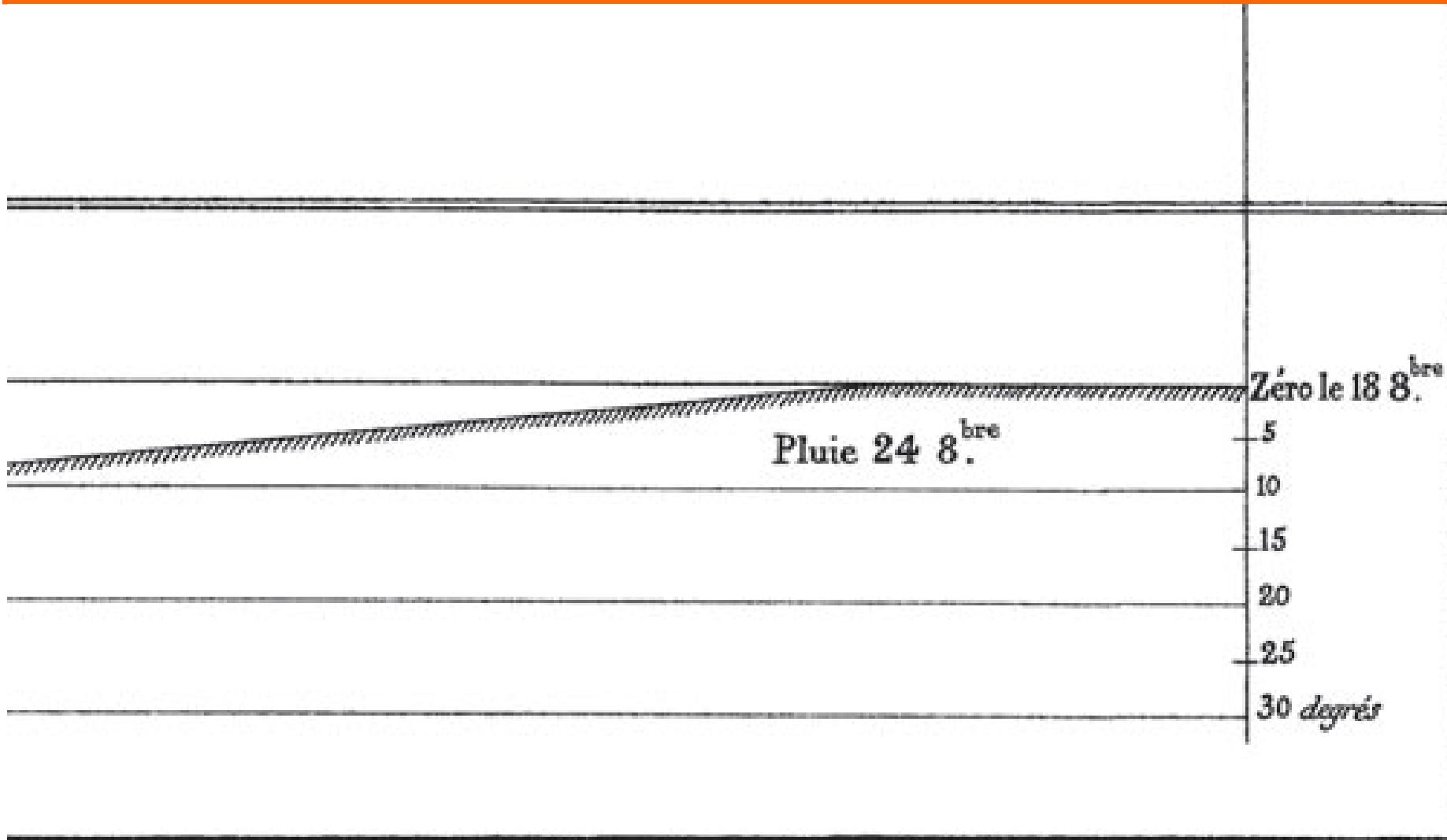
Imp. Lith. Regnier et Dureuil.

https://en.wikipedia.org/wiki/Charles_Joseph_Minard

Numbers and direction

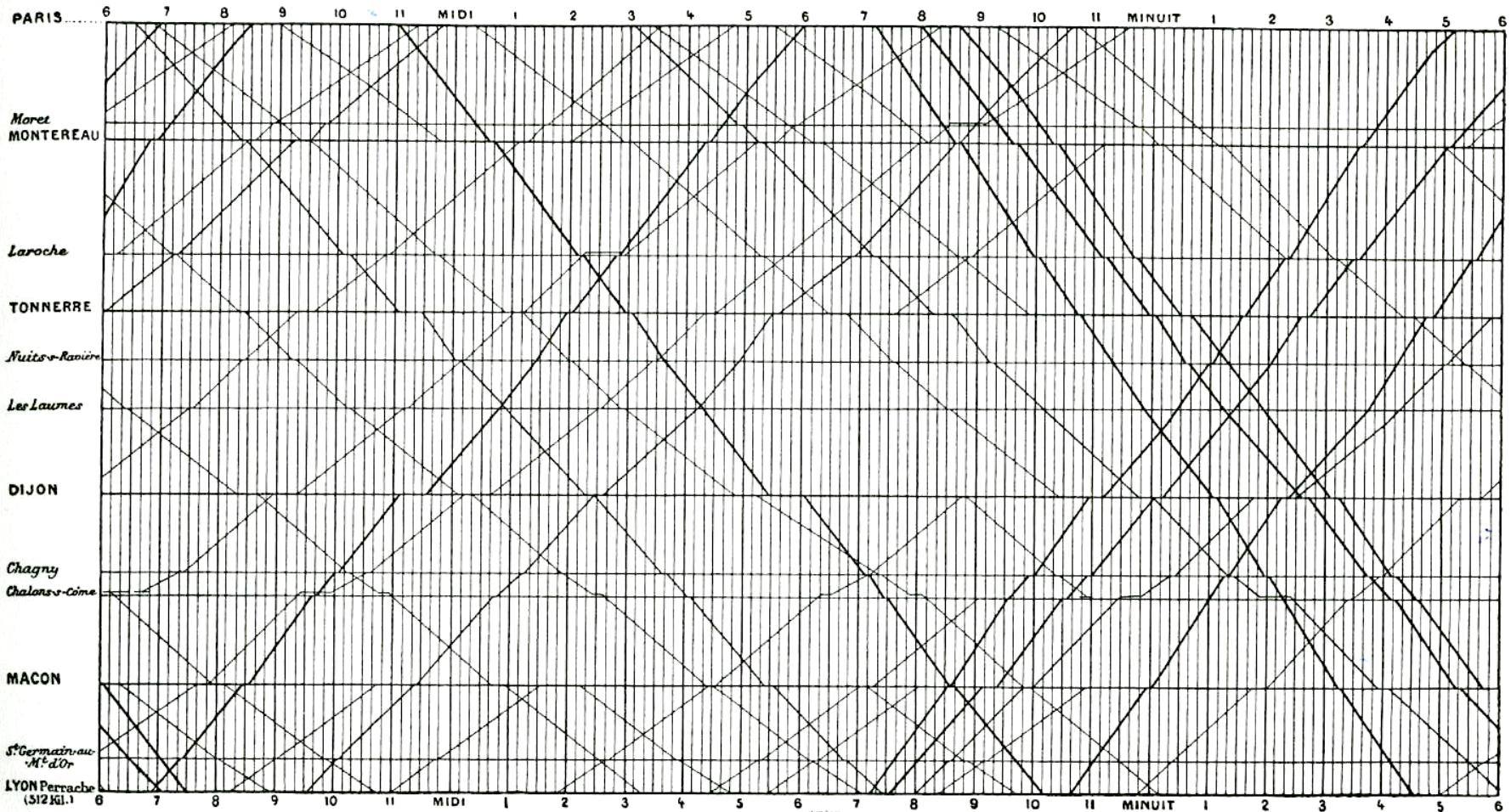


Temperature



Imp. Lith. Regnier et Dourdet.

Étienne-Jules Marey



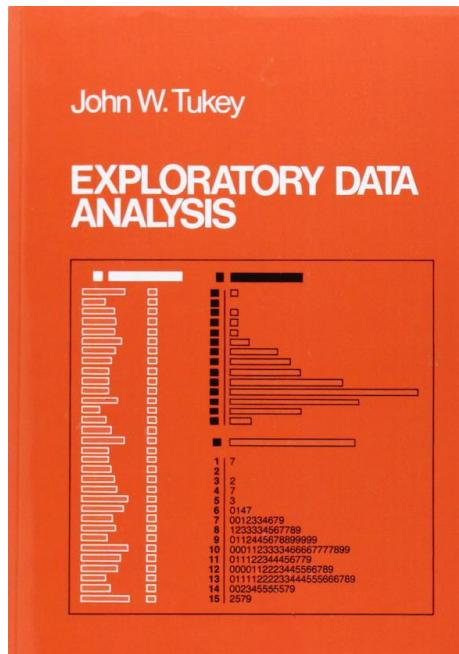
La Méthode graphique dans les sciences expérimentales et principalement en physiologie et en médecine, 1885

<https://archive.org/details/lamthodegraphiq00maregoog>

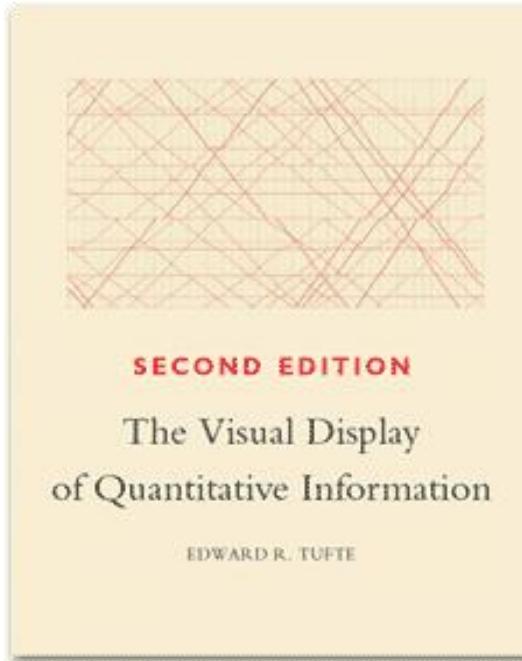
XX Century

- <http://www.datavis.ca/milestones/>

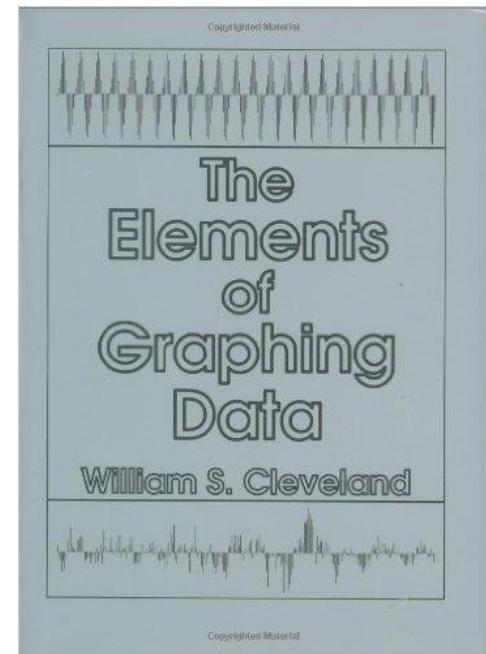
1977



1983



1985



INFORMATION VISUALIZATION

Information visualization

The use of computer-supported,
interactive, visual representations
of abstract data to amplify
cognition

Readings in Information Visualization: Using Vision to Think.
S.K.Card, J.D.Mackinlay, and B.Shneiderman, Academic Press, 1999

Overview

Understanding → Decisions

Information Visualization

Visual Patterns, Trends, Exceptions

Quantitative Reasoning

Quantitative Relationship & Comparison

Visual Perception

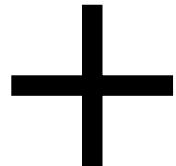
Visual Properties & Objects

Data

Representation/Encoding

Quantitative message

- Quantitative values
 - ◆ Express measures



- Categories
 - ◆ Identify what entities the values refer to
 - ◆ Define groups of entities

Understanding tasks

- Variation within quantitative measures
 - ◆ Distribution
 - ◆ Deviation
 - ◆ Correlation
- Variation within category
 - ◆ Ranking
 - ◆ Part-to-whole
 - ◆ Time
 - ◆ Space
- Multivariate

Visualization instruments

- Tables
 - ◆ Textual information
- Graphs
 - ◆ Visual information

Tables

- Main features
 - ◆ Data arranged in rows and columns
 - ◆ Data encoded as **text**
- Strengths
 - ◆ Easy **look-up** of values
 - ◆ Precise values
 - Allow selected comparisons
 - ◆ Several units of measure are possible

Graphs

- Main features
 - ◆ One or more **axes** delineate the display area where values are shown
 - ◆ Values encoded as **visual** objects in relation to axes
 - ◆ Axes provide **scales**
 - Assign values and labels to visual objects
 - Both categorical and quantitative
- Strengths
 - ◆ Overall shape of data (holistic)

Graphs

- Show
 - ◆ Trend
 - Pattern of change over time
 - ◆ Comparison of subsets
 - Overall
 - Spot similarities and differences
 - ◆ Highlight exceptions
- Display relationships among multiple quantitative values by giving them shape

In general

Use tables to

Look up individual values

Compare individual values

Precise values are required

There is more than one unit of measure

Use graphs to

Focus on the shape of values

Reveal relationships among multiple values

EXAMPLES

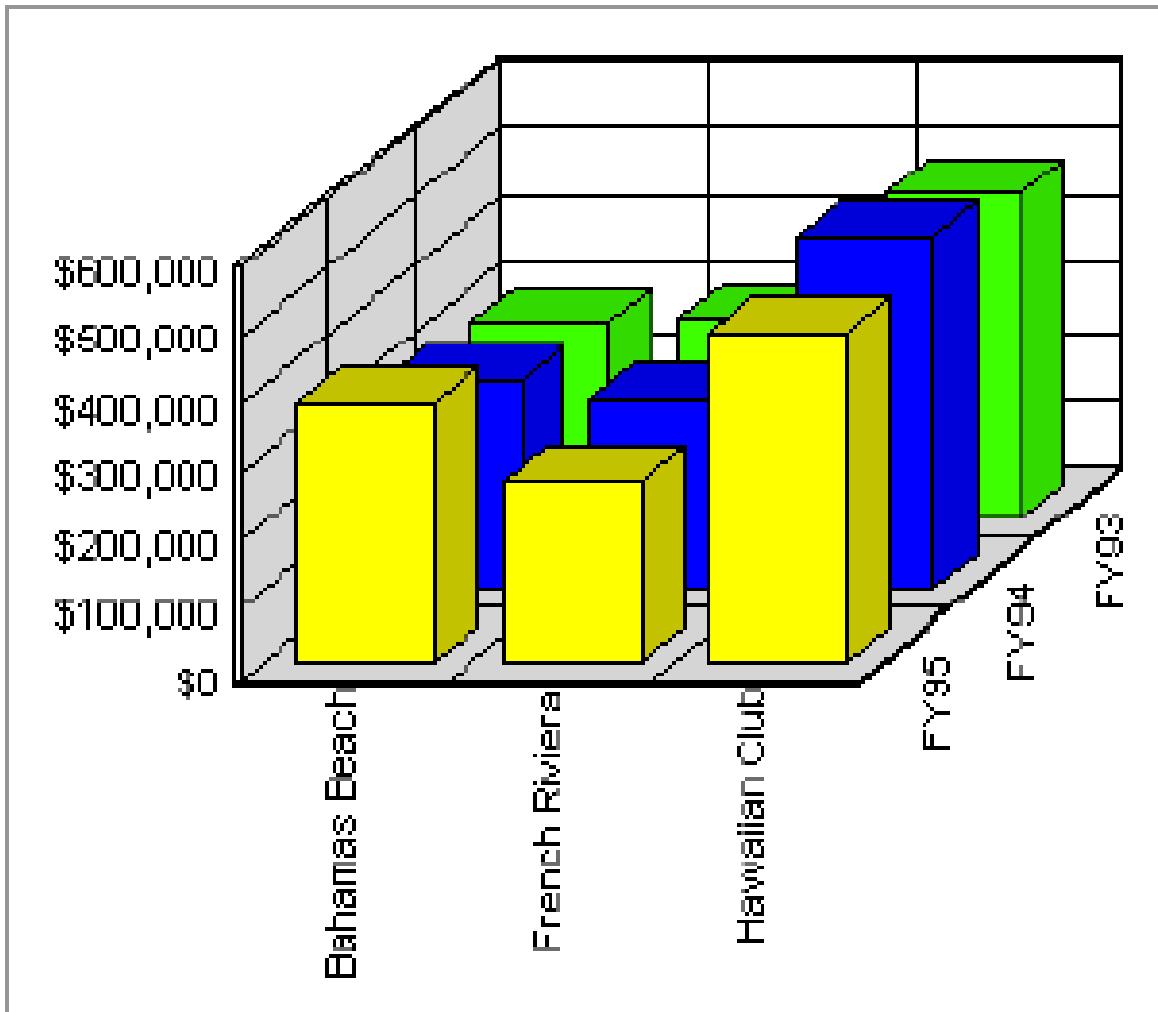
Good and Poor visualization

- Like good writing, good graphical displays of data communicate ideas with clarity, precision, and efficiency.
- Like poor writing, bad graphical displays distort or obscure the data, make it harder to understand or compare, or otherwise thwart the communicative effect which the graph should convey.

Friendly, Michael, and Daniel J. Denis. (2001)

"Milestones in the history of thematic cartography, statistical graphics, and data visualization."
<http://www.datavis.ca/milestones>

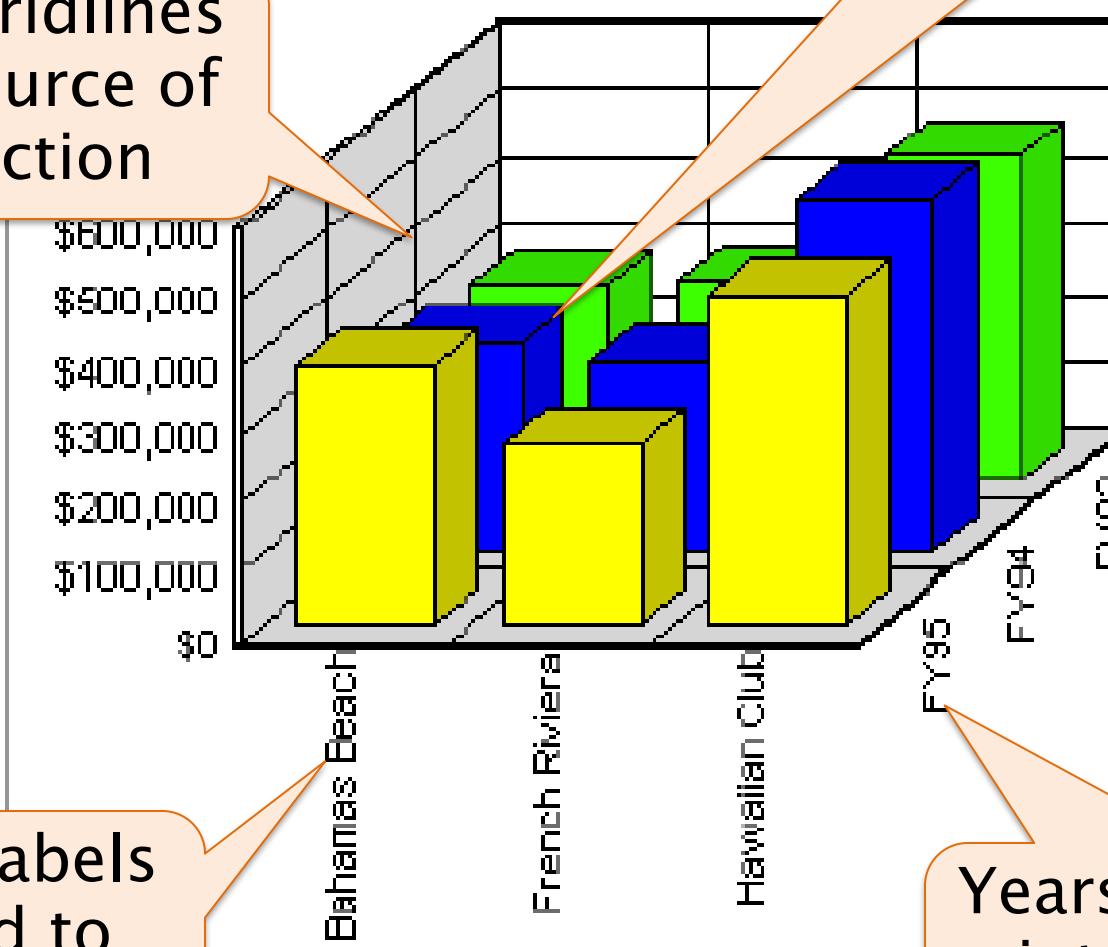
A bar graph



A **bad** bar graph

3D bars are impossible to read

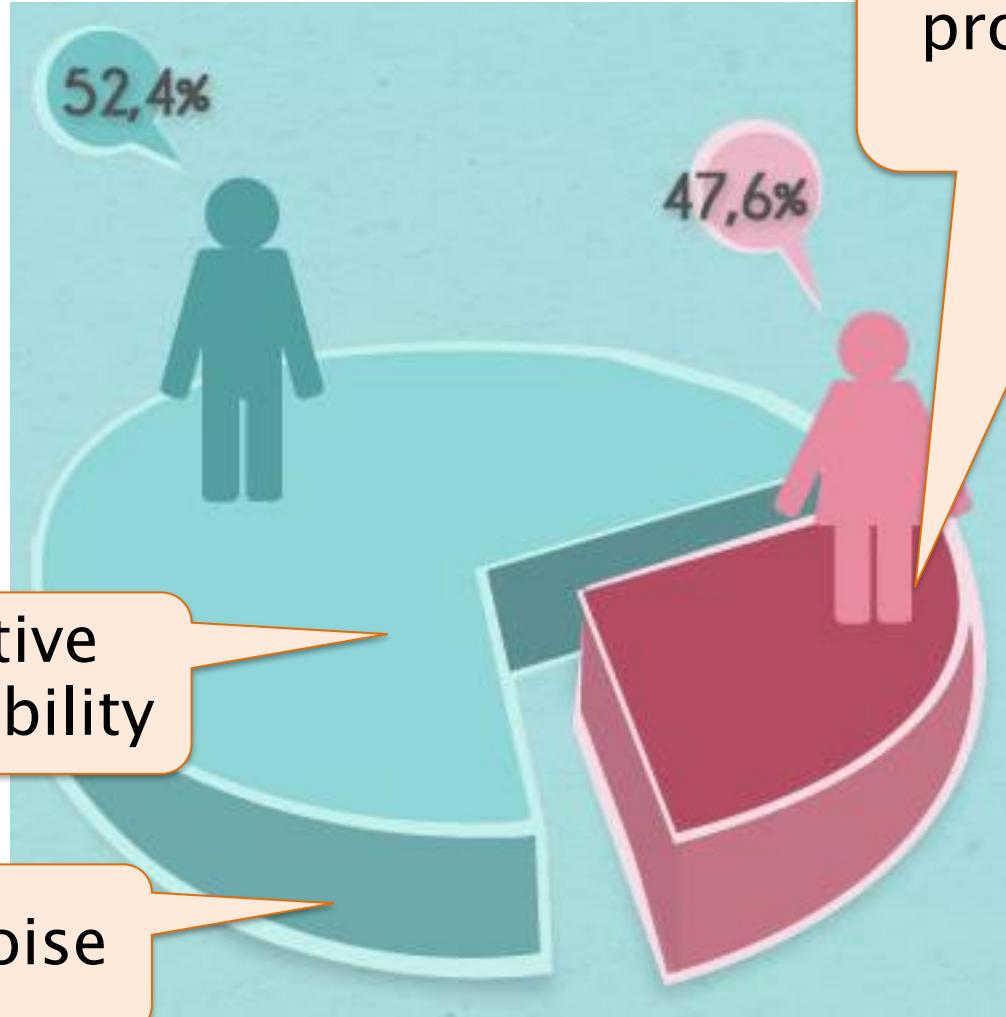
Heavy gridlines are a source of distraction



Vertical labels are hard to read

Years run counter-intuitively from back to front

A pie chart



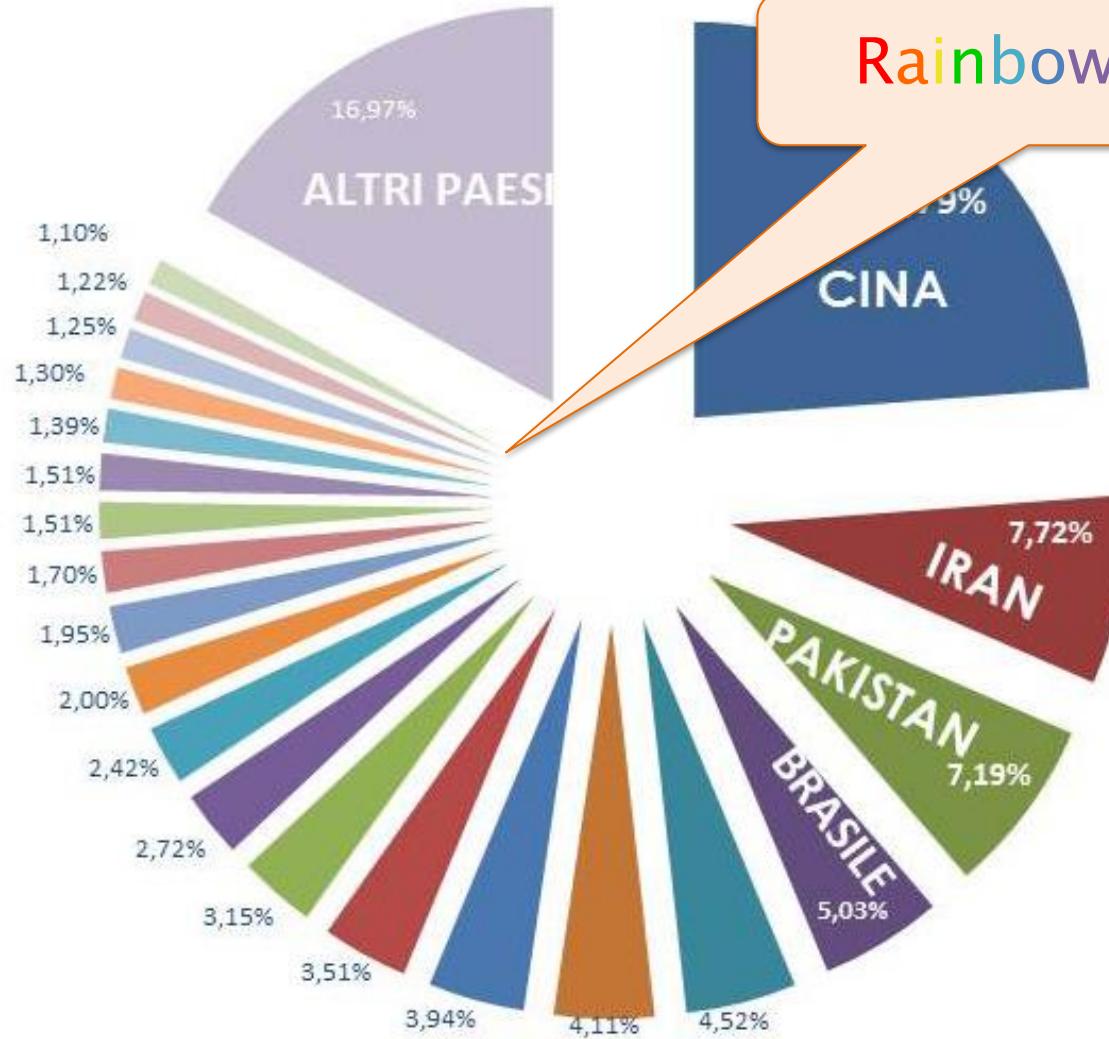
3D Perspective
worsen readability

Depth add noise

Size of slices not
proportional to
values

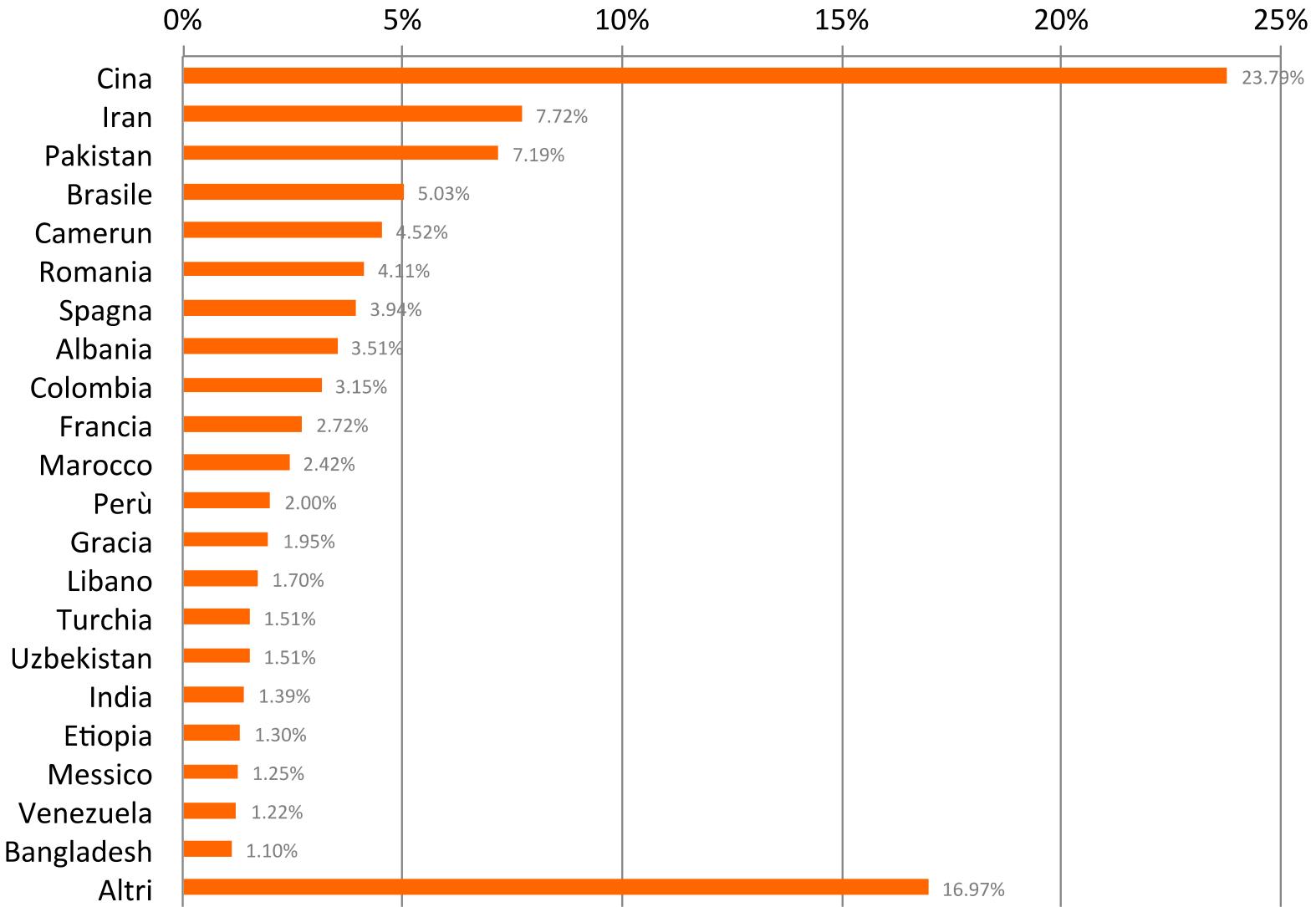
Pie chart (original)

- CINA
- IRAN
- PAKISTAN
- BRASILE
- CAMERUN
- ROMANIA
- SPAGNA
- ALBANIA
- COLOMBIA
- FRANCIA
- MAROCCO
- PERU'
- GRECIA
- LIBANO
- TURCHIA
- UZBEKISTAN
- INDIA
- ETIOPIA
- MESSICO
- VENEZUELA
- BANGLADESH
- ALTRI PAESI

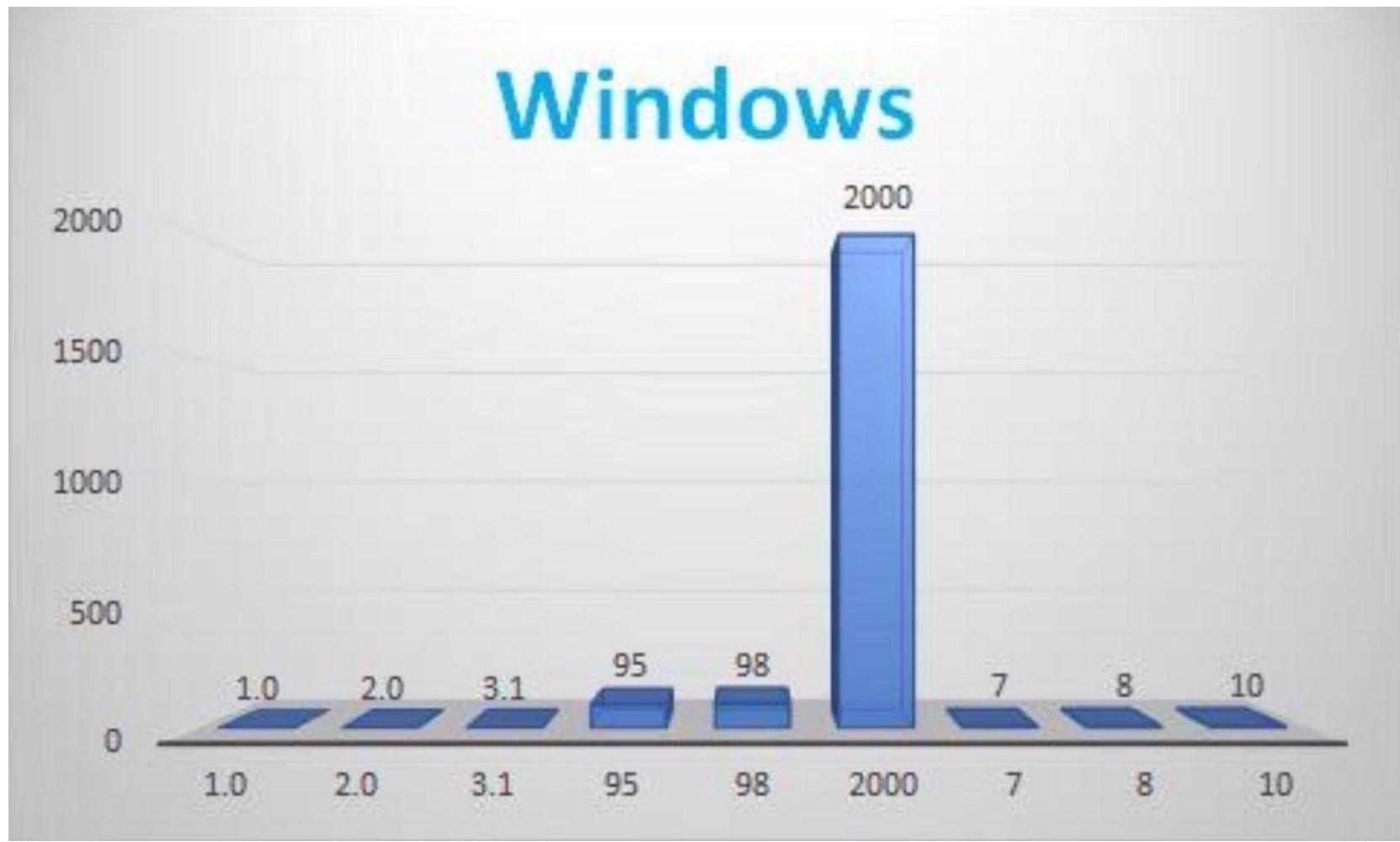


Rainbow Effect

Bar chart (redesign)



Meaningless Data



The David Bowie Song That Fans Are Listening to Most: 'Heroes'

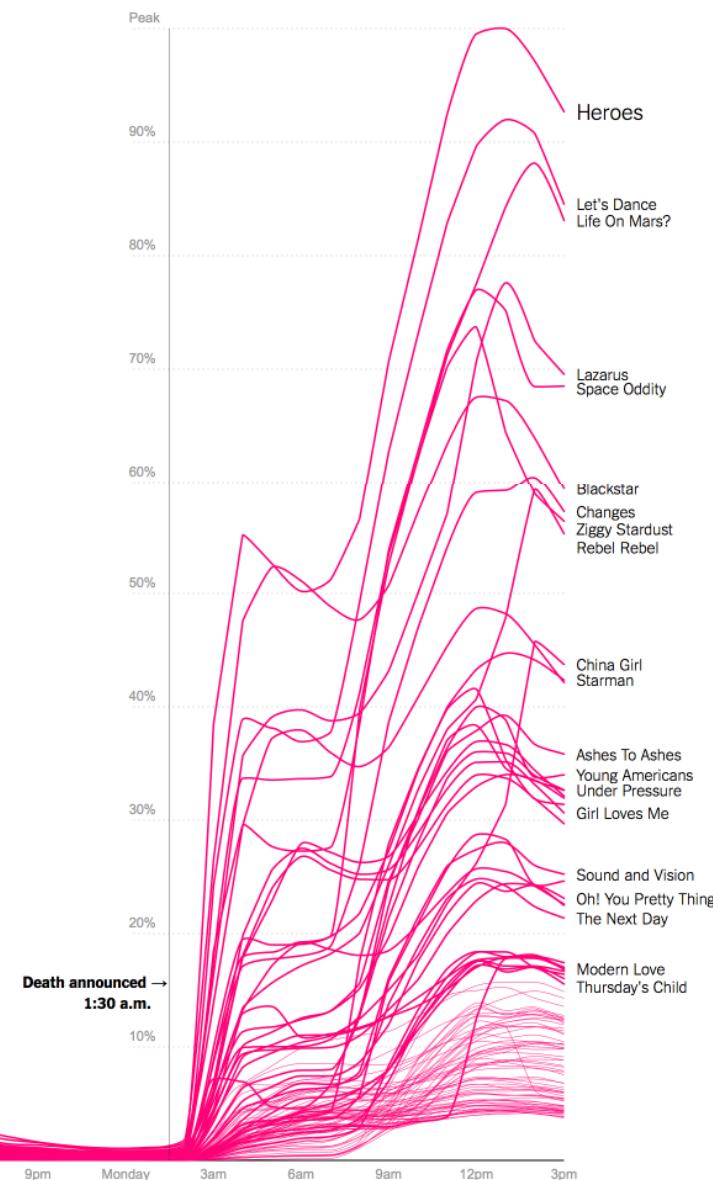
By QUOCTRUNG BUI, JOSH KATZ and JASMINE C. LEE JAN. 12, 2016

Popularity of David Bowie songs on Spotify

Plays per hour

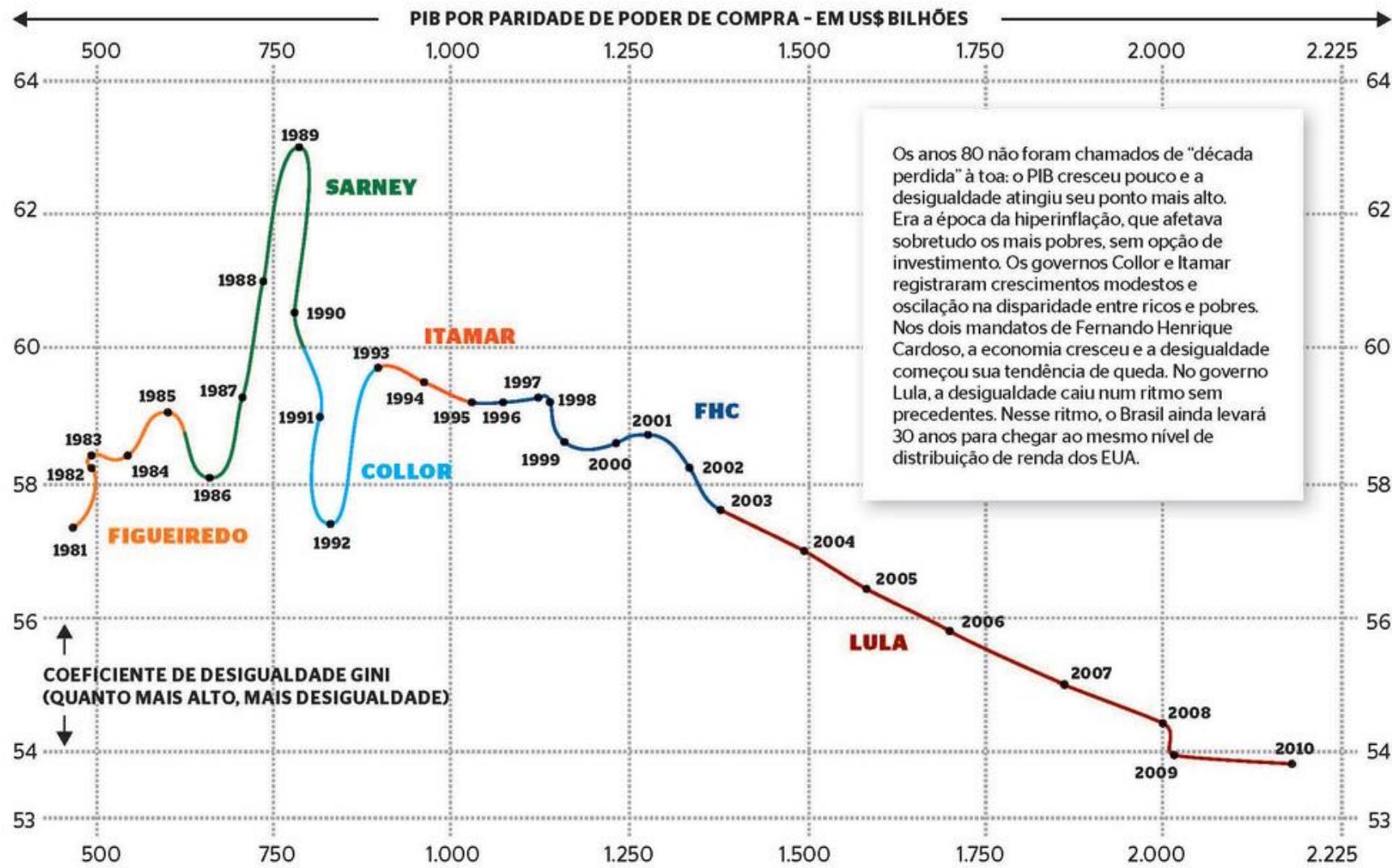


Plays are expressed as a percentage of the maximum hourly plays for the leading song.



Quando o PIB cresce, nem sempre a desigualdade cai

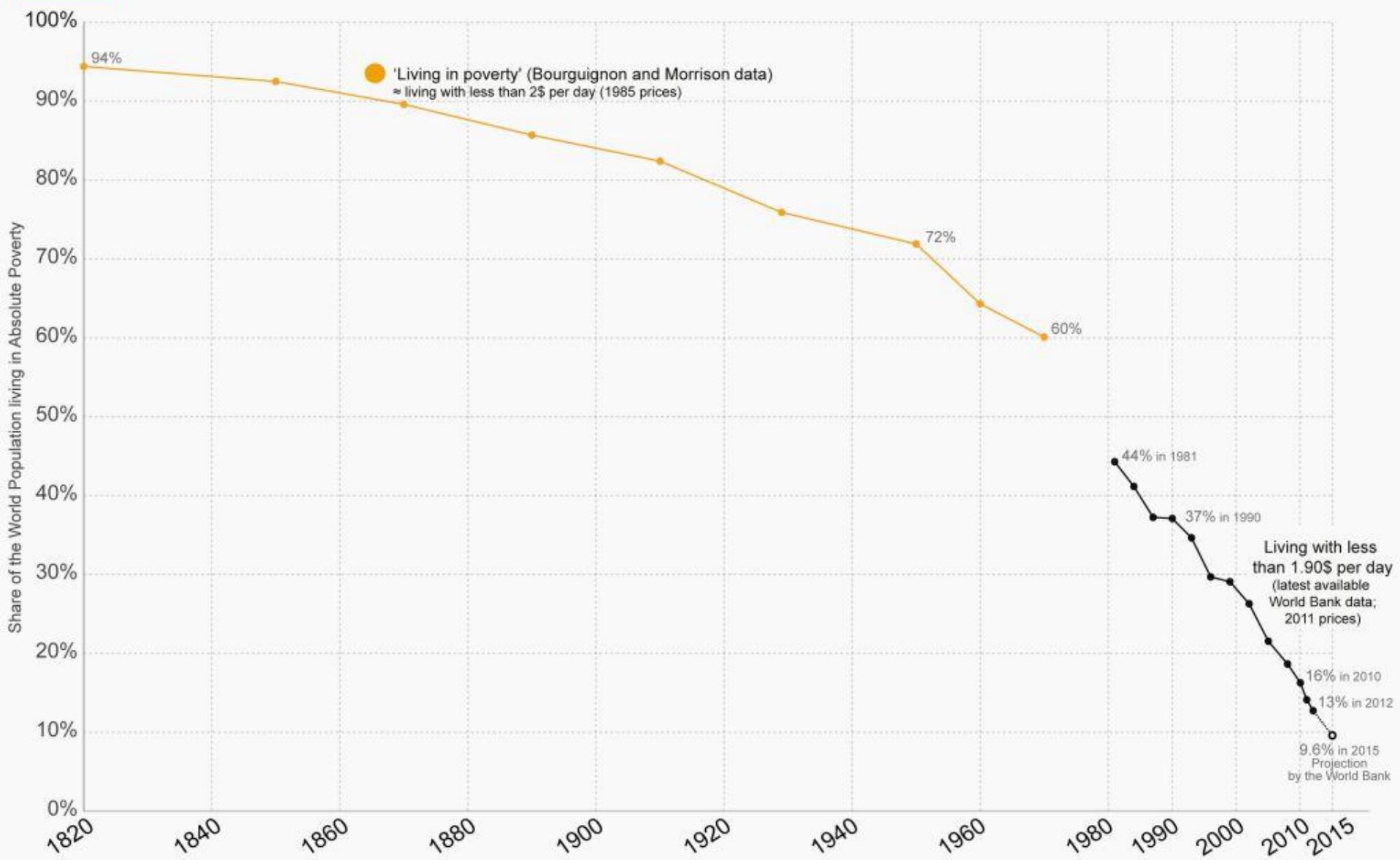
O gráfico abaixo mostra o avanço do PIB comparado à evolução da desigualdade no Brasil desde 1980. Nem sempre o crescimento econômico levou a uma redução proporcional na disparidade de renda entre os mais pobres e os mais ricos



Fontes: Banco Mundial, FMI, IBGE Gráficos : Alberto Cairo

Share of the World Population living in Absolute Poverty, 1820-2015 – by Max Roser

All incomes are adjusted for inflation over time and for price differences between countries (1985-PPP before 1970; 2011-PPP after 1970).



Data sources: 1820-1970 Bourguignon and Morrison (2002) - Inequality among World Citizens, In The American Economic Review; 1981-2015 World Bank (PovcalNet)

The interactive data visualisation is available at OurWorldInData.org. There you find the raw data and more visualisations on this topic.

Licensed under CC-BY-SA by the author Max Roser.

Migrants arrived in period January – June

2019

2018

2017

The accidents at work happened and reported

Migrants arrived in period January – June

2019

2018

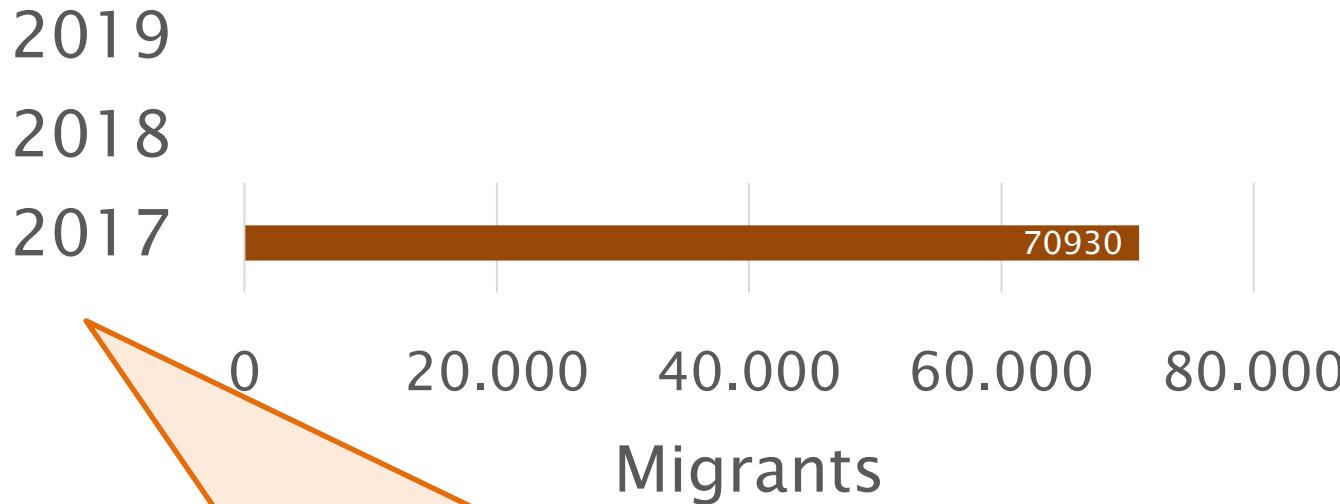
2017



What is the order of magnitude
of migrants arrived in 2017?

1k, 10k, 20k, 40k, 80k

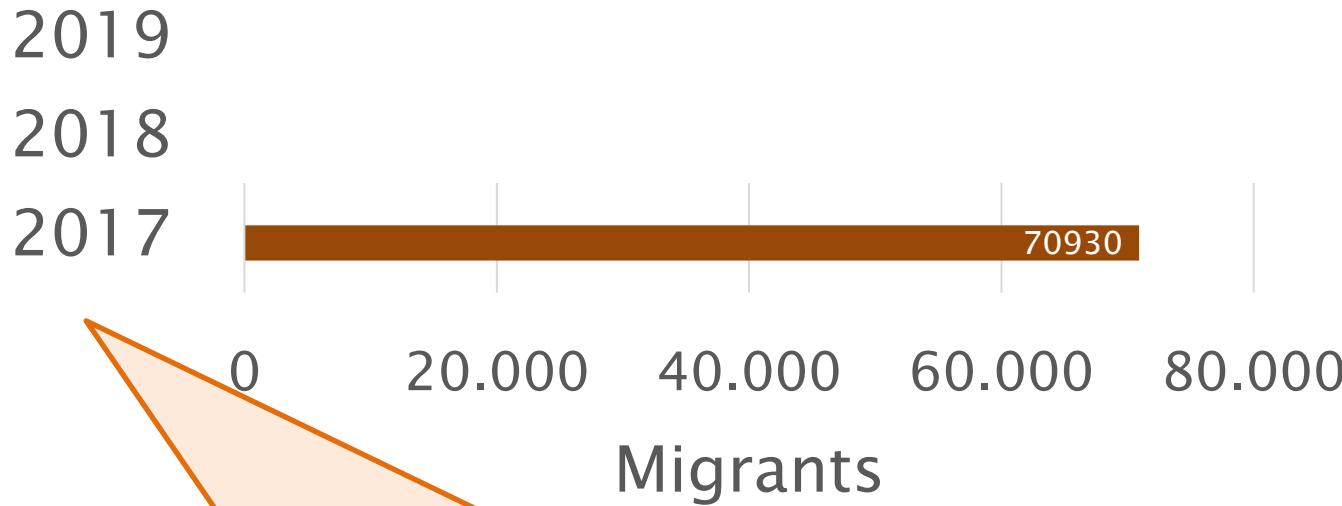
Migrants arrived in period January – June



What is the order of magnitude of migrants arrived in 2017?

1k, 10k, 20k, 40k, 80k

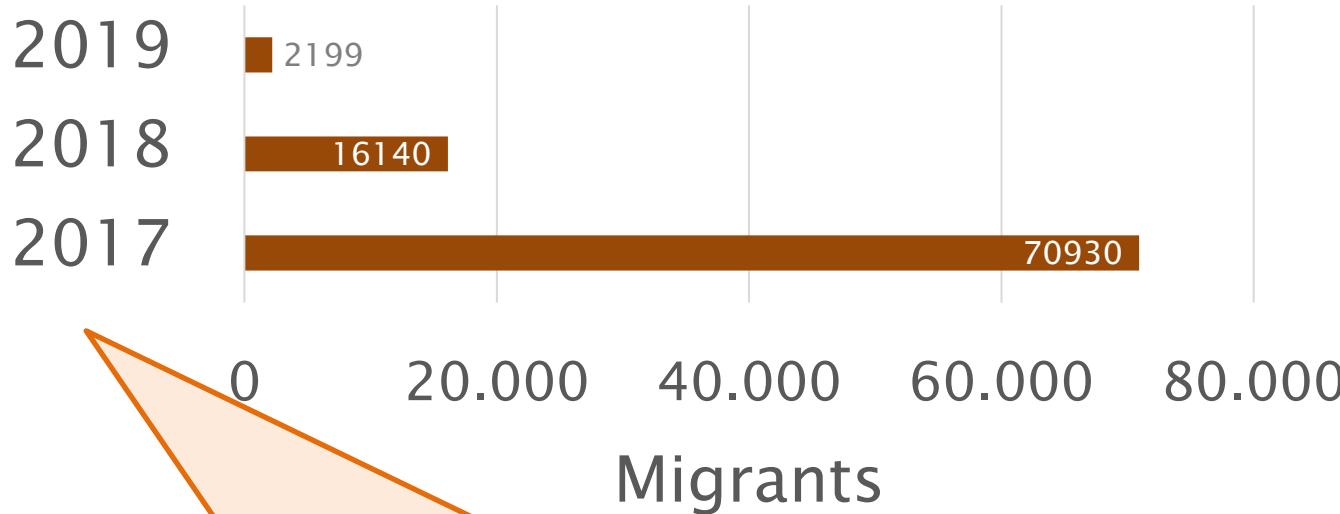
Migrants arrived in period January – June



The ratio of number of migrants in 2018, with respect to 2017 is

1:1 , 1:2 , 1:4 , 1:10 , 1:20

Migrants arrived in period January – June



The ratio of number of migrants in 2018, with respect to 2017 is

1:1 , 1:2 , 1:4 , 1:10 , 1:20

What is the order of magnitude
of accidents in Q1 2019?

1k, 50k, 100k, 200k, 500k

The **accidents at work** happened and reported to Inail in first quarter 2019 have been

What is the order of magnitude
of accidents in Q1 2019?

1k, 50k, **100k**, 200k, 500k

The **accidents at work** happened and reported to Inail in first quarter 2019 have been 131 thousand (109 thousand at work and 22 thousand while traveling),

With respect to Q1 2018 how much have changed accidents in Q1 2019?

-5k , -2k , ±500 , +2k , +5k

The **accidents at work** happened and reported to Inail in first quarter 2019 have been 131 thousand (109 thousand at work and 22 thousand while traveling),

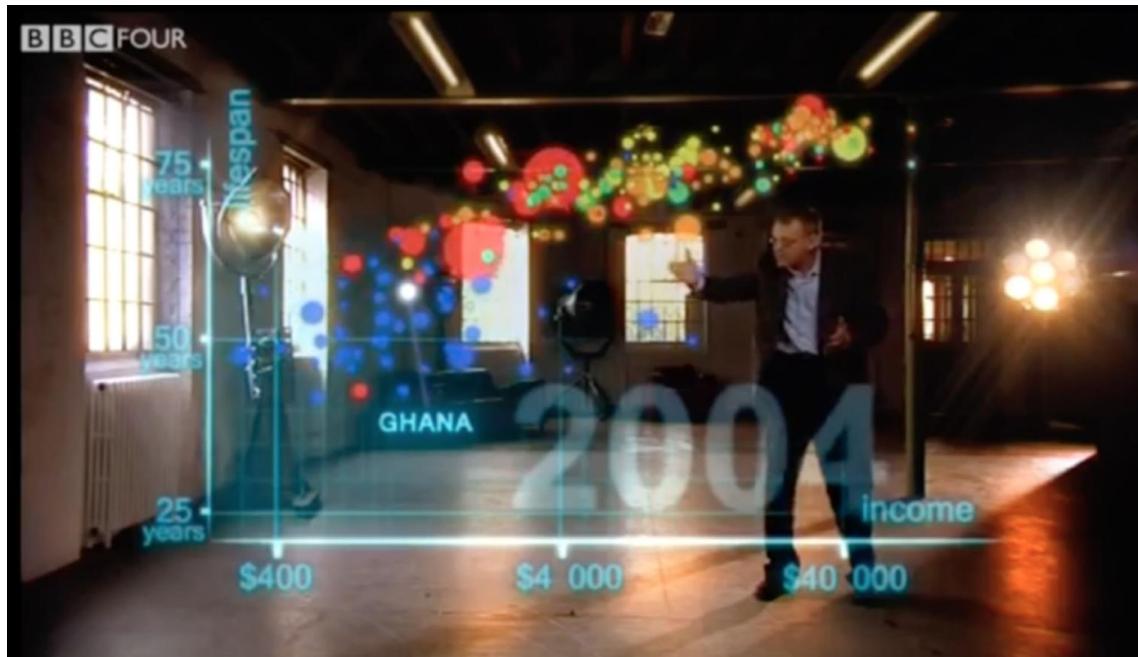
With respect to Q1 2018 how much have changed accidents in Q1 2019?

-5k , -2k , ±500 , +2k , +5k

The **accidents at work** happened and reported to Inail in first quarter 2019 have been 131 thousand (109 thousand at work and 22 thousand while traveling), increased by 1.7% (+2 thousand reports) with respect to first quarter 2018

Hans Rosling (1948-2017)

- 200 Countries, 200 Years, 4 Minutes
 - ◆ The Joy of Stats – BBC 4
 - <http://www.bbc.co.uk/programmes/b00wgq0l>
 - <https://www.youtube.com/watch?v=jbkSRLYSojo>



References

- The picture superiority effect in associative recognition.
 - ◆ <https://www.ncbi.nlm.nih.gov/pubmed/18927048>
- John W. Tukey, 1977. Exploratory Data Analysis, Pearson.
- Edward R. Tufte, 1983. The Visual Display of Quantitative Information. Graphics Press.
- William S. Cleveland, 1994,
The Elements of Graphing Data, Hobart Press
- S.K.Card, J.D.Mackinlay, and B.Shneiderman.
Readings in Information Visualization: Using Vision to Think. Academic Press, 1999

Visual perception

Data Management and Visualization



SoftEng
<http://softeng.polito.it>

Version 2.2.1
© Marco Torchiano, 2021





This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

To view a copy of this license, visit

<http://creativecommons.org/licenses/by-nc-nd/4.0/>.

You are free: to copy, distribute, display, and perform the work

Under the following conditions:



Attribution. You must attribute the work in the manner specified by the author or licensor.



Non-commercial. You may not use this work for commercial purposes.



No Derivative Works. You may not alter, transform, or build upon this work.

- For any reuse or distribution, you must make clear to others the license terms of this work.
- Any of these conditions can be waived if you get permission from the copyright holder.

Your fair use and other rights are in no way affected by the above.

VISUAL INTEGRITY

Principles of integrity

- **Proportionality**
 - ◆ Representation as physical quantities should be proportional to the represented numbers
- **Utility**
 - ◆ Graphical element should convey useful information
- **Clarity**
 - ◆ Labeling should counter graphical distortion and ambiguity

Proportionality

- The magnitude of visual attributes should represent faithfully the magnitude of measures
- They should allow
 - ◆ Discrimination: are they different?
 - ◆ Comparison: which is larger?
 - ◆ Magnitude Assessment: how much larger?

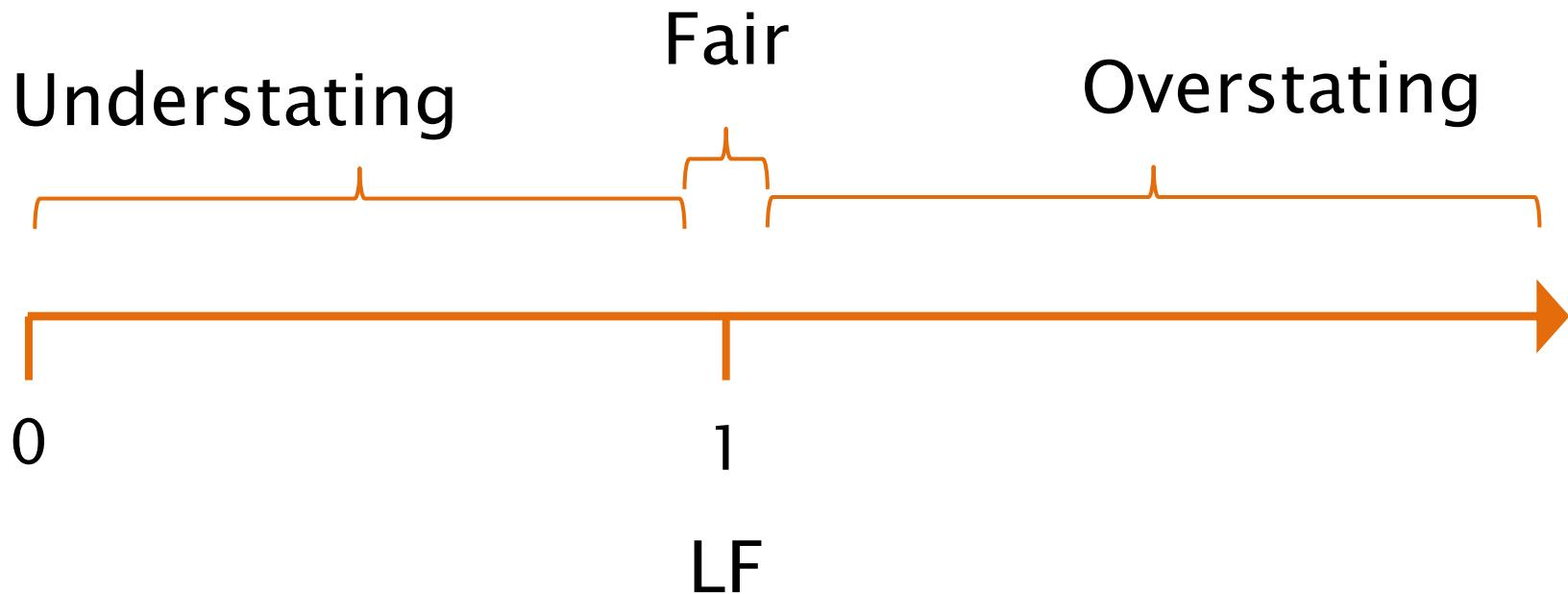
Lie Factor

$$LF = \frac{\text{size of effect shown in graphic}}{\text{size of effect in data}}$$

- Overstating
 - ◆ $LF > 1 \Leftrightarrow \log(LF) > 0$
- Understating
 - ◆ $LF < 1 \Leftrightarrow \log(LF) < 0$
- Fair
 - $LF = 1 \Leftrightarrow \log(LF) = 0$

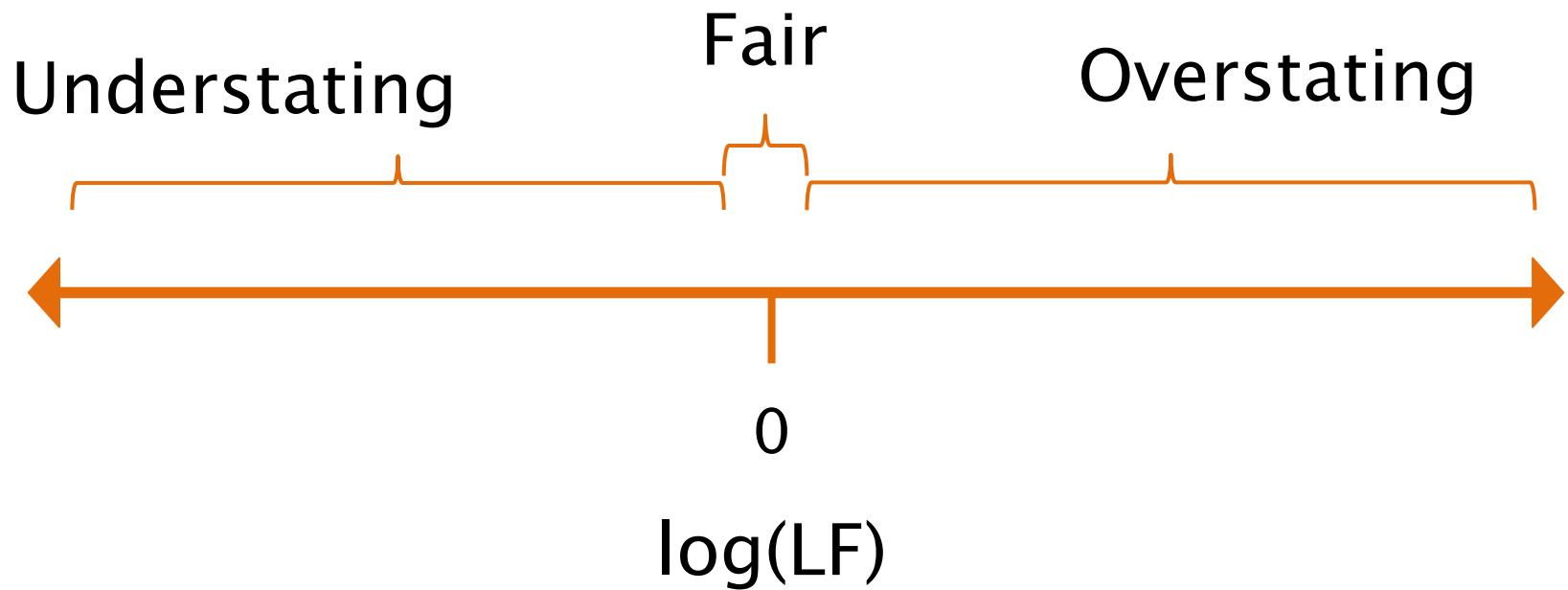
Lie Factor

$$LF = \frac{\text{size of effect shown in graphic}}{\text{size of effect in data}}$$

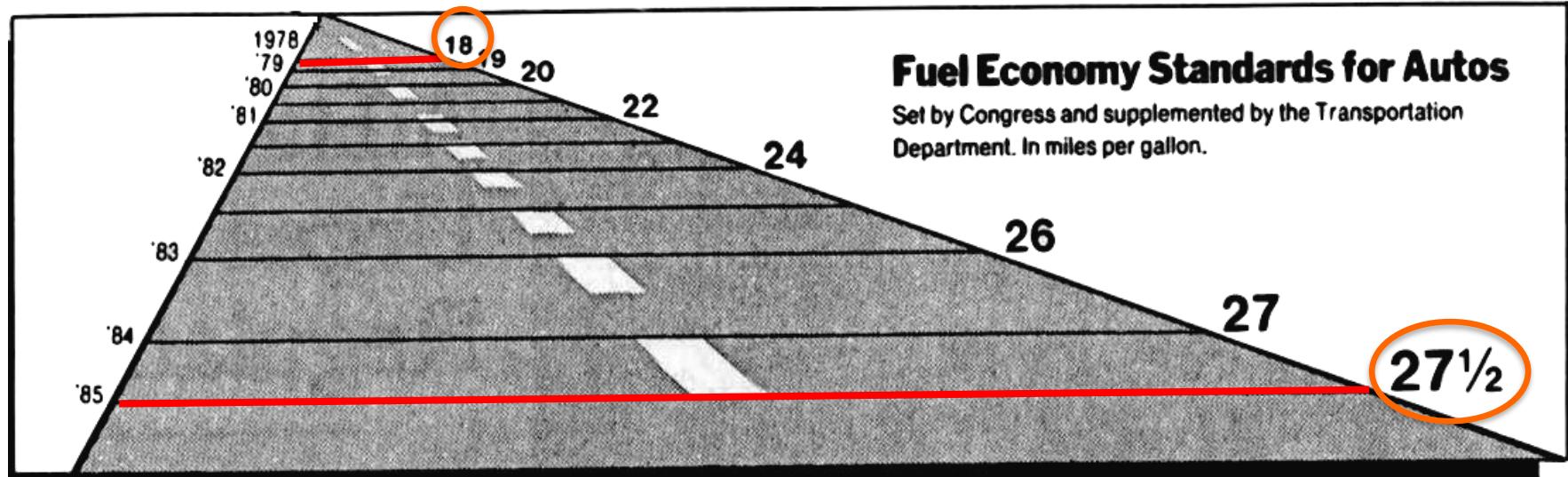


Lie Factor

$$LF = \frac{\text{size of effect shown in graphic}}{\text{size of effect in data}}$$



Lie Factor



$$\frac{18.7}{2.2} = 8.5 \text{ on graphic}$$

$$\frac{27.5}{18} = 1.52 \text{ in data}$$

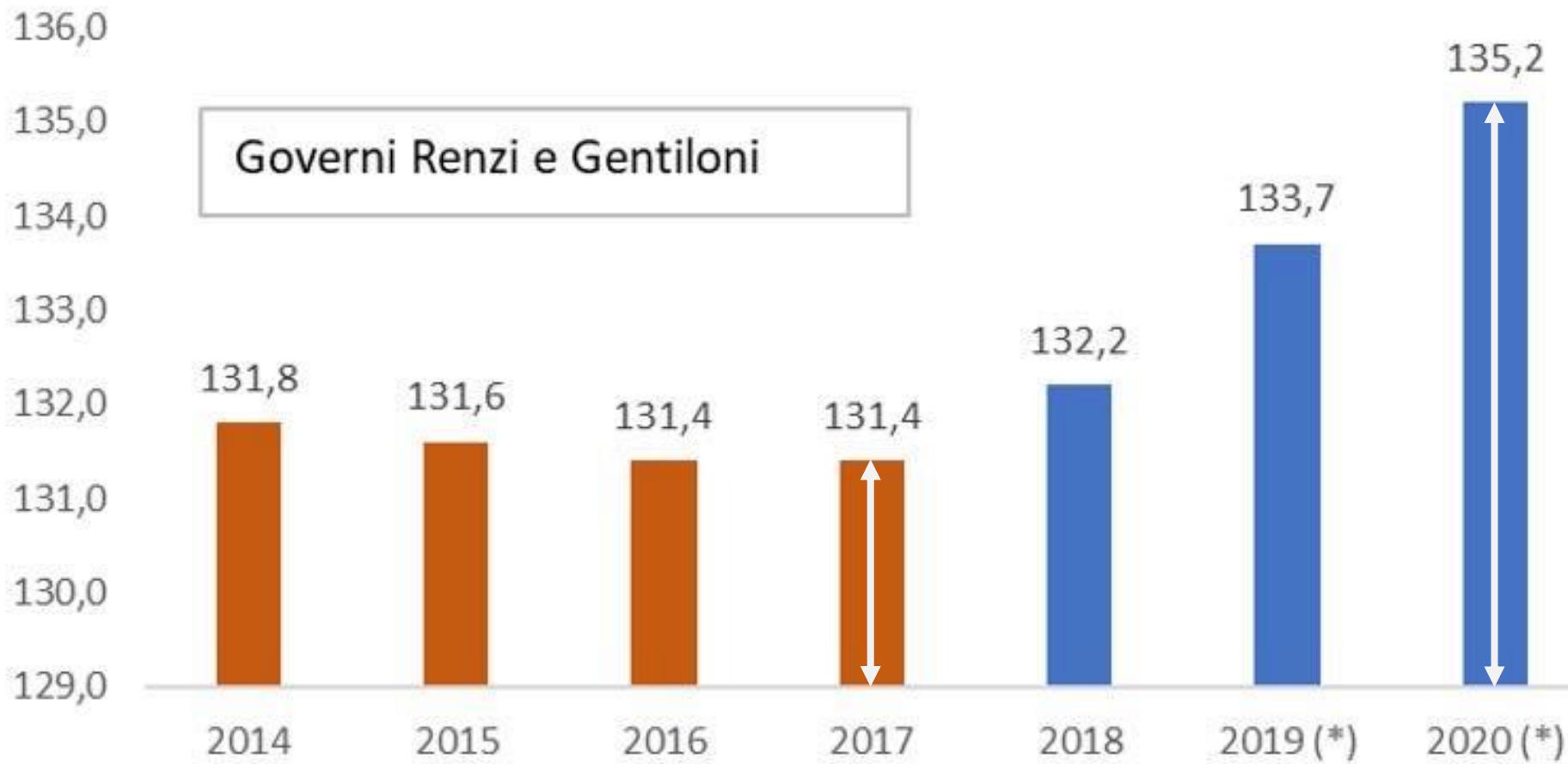
$$LF = 8.5 / 1.52 = 5.59$$

Example

Debito pubblico (% PIL)

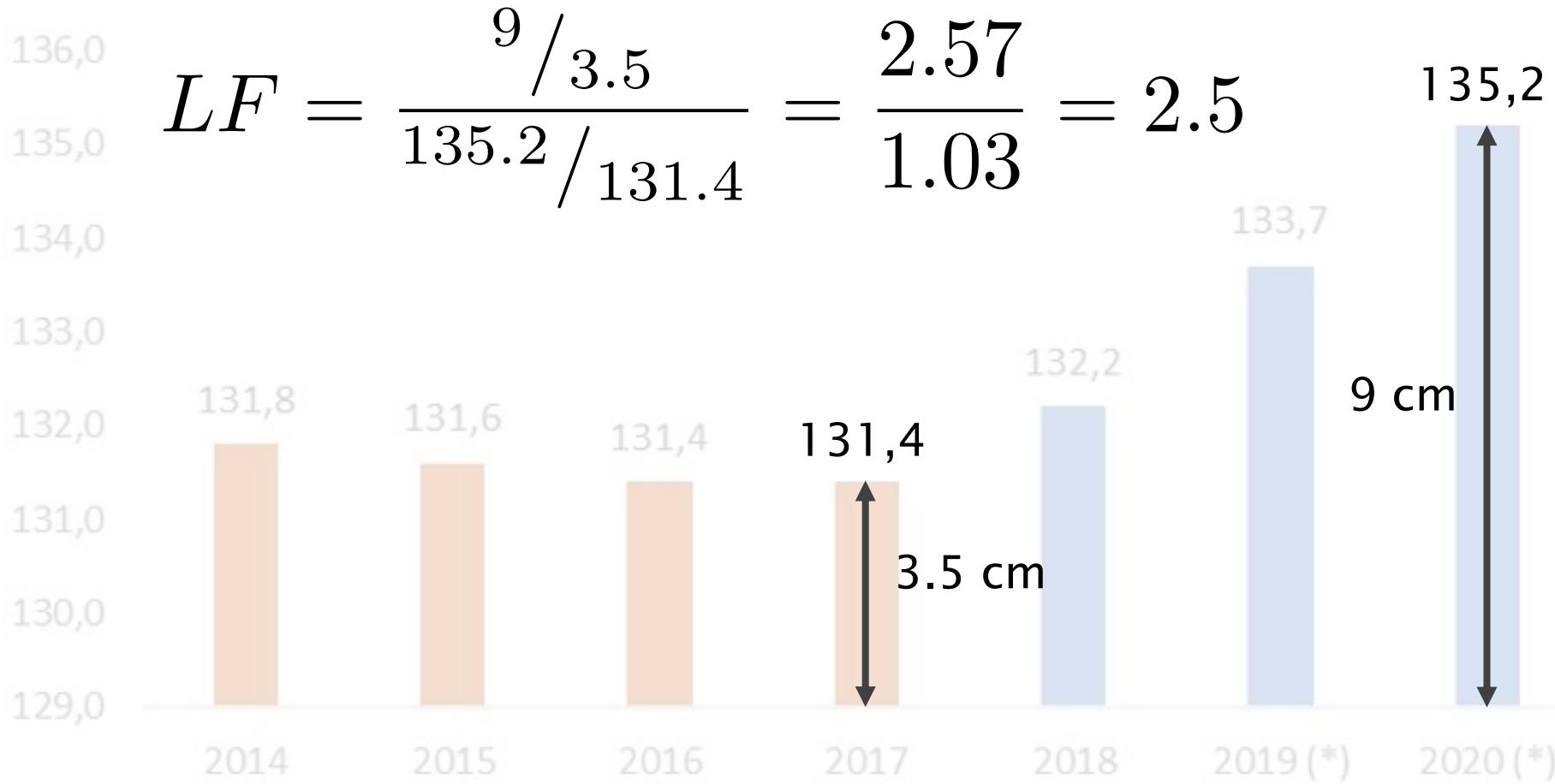
(*) previsioni Commissione UE

Governo Conte



Example - Lie Factor

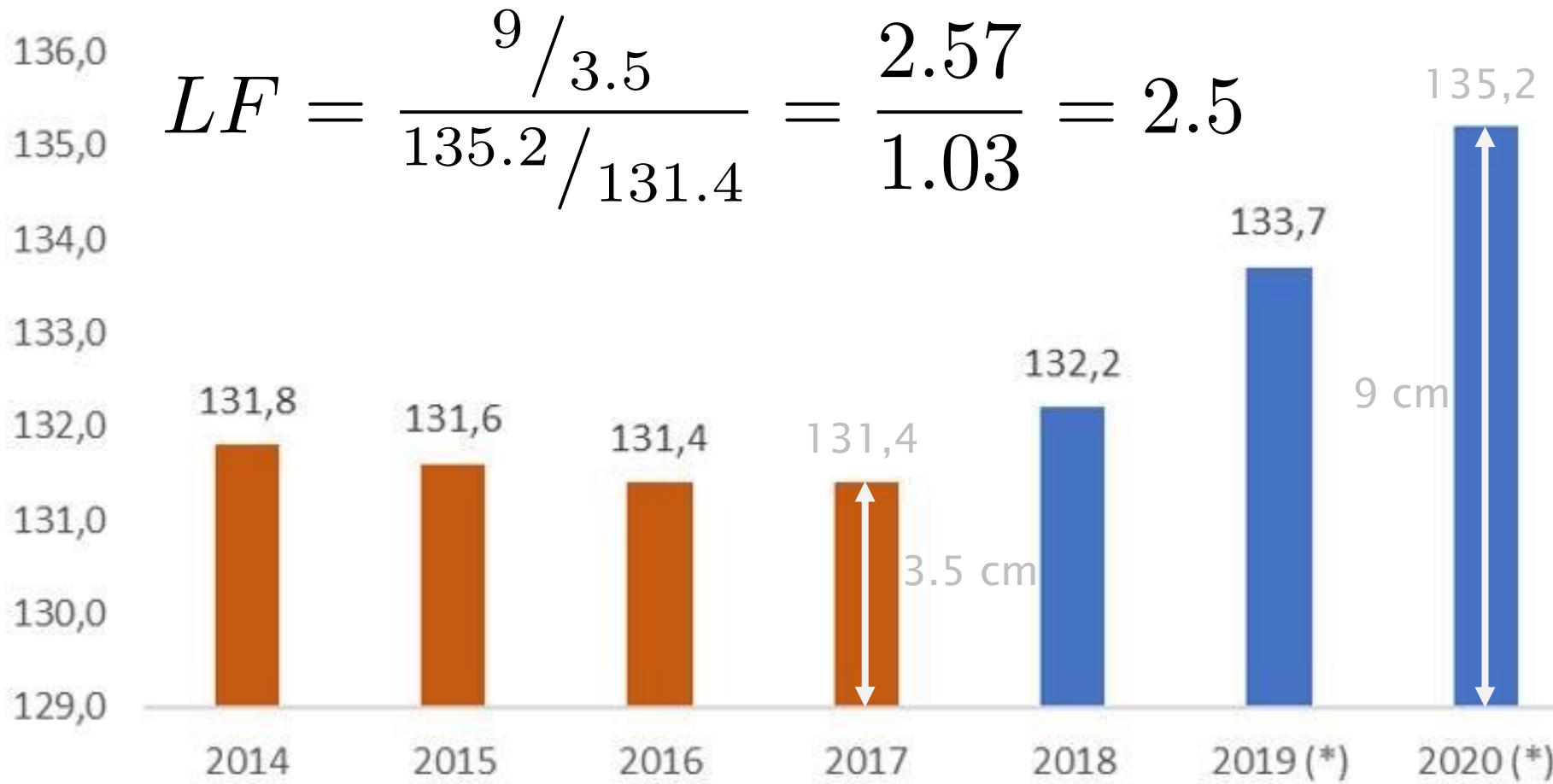
Debito pubblico (% PIL)
(*) previsioni Commissione UE



Example - Lie Factor

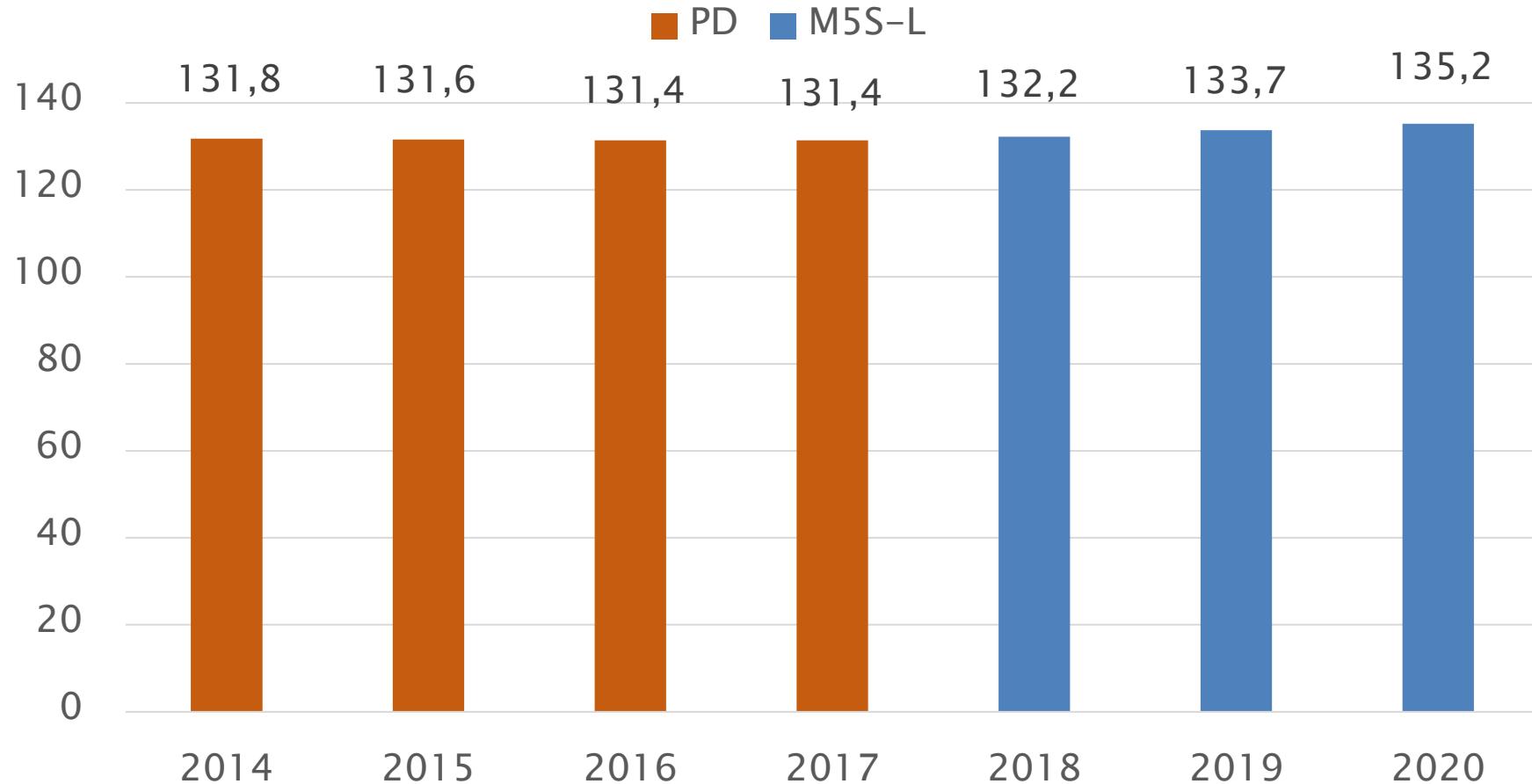
Debito pubblico (% PIL)

(*) previsioni Commissione UE



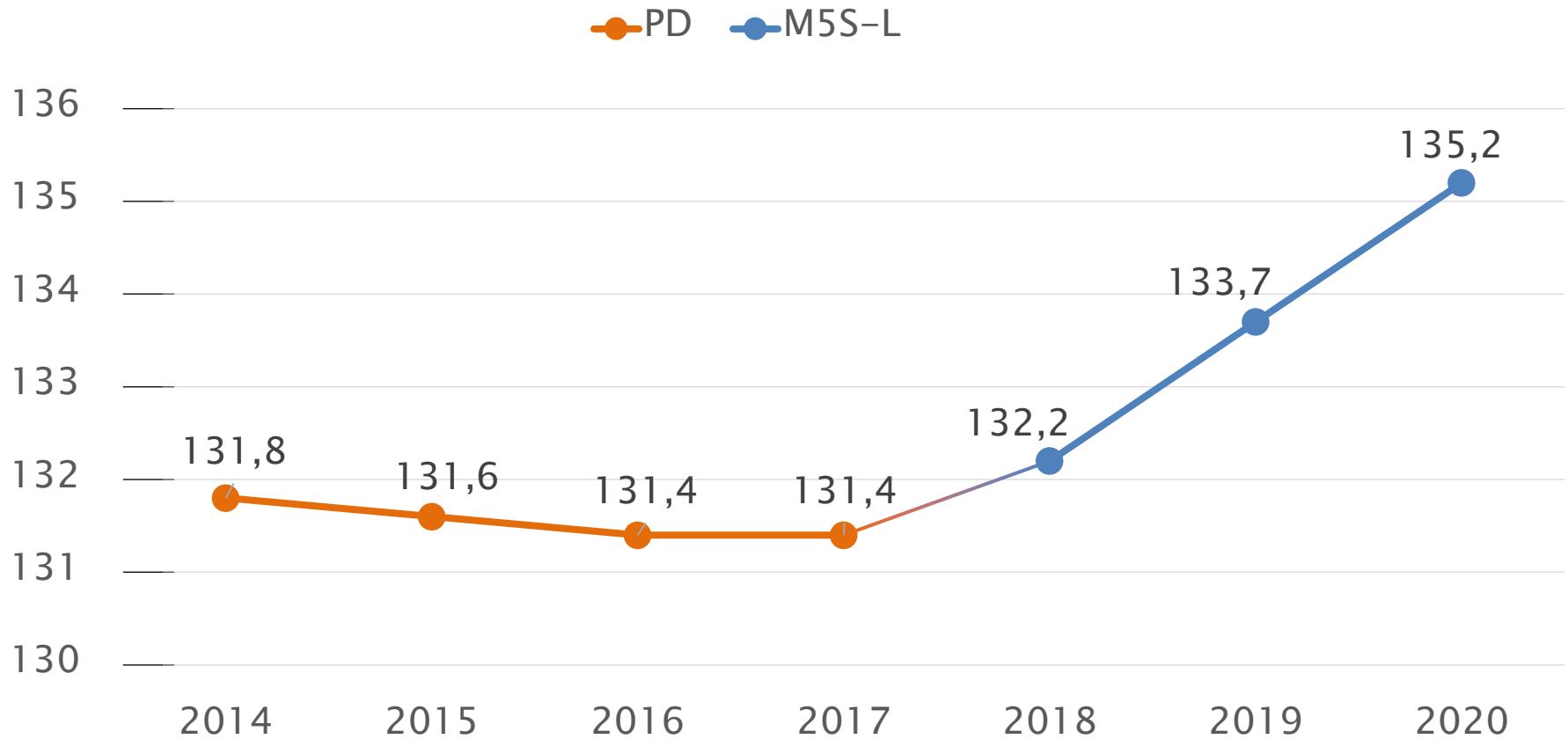
Example – Redesign

Debito Pubblico (% PIL)



Example – Redesign

Debito Pubblico (% PIL)



Guidelines for design

- Keep the physical Lie Factor = 1
- Limit the perceptual Lie Factor as much as possible

Utility

- Every element should convey useful information
- Unnecessary visual objects or attributes distract from the message
 - ◆ Different attributes trigger a search for a rationale (e.g. random colors)

Data-ink

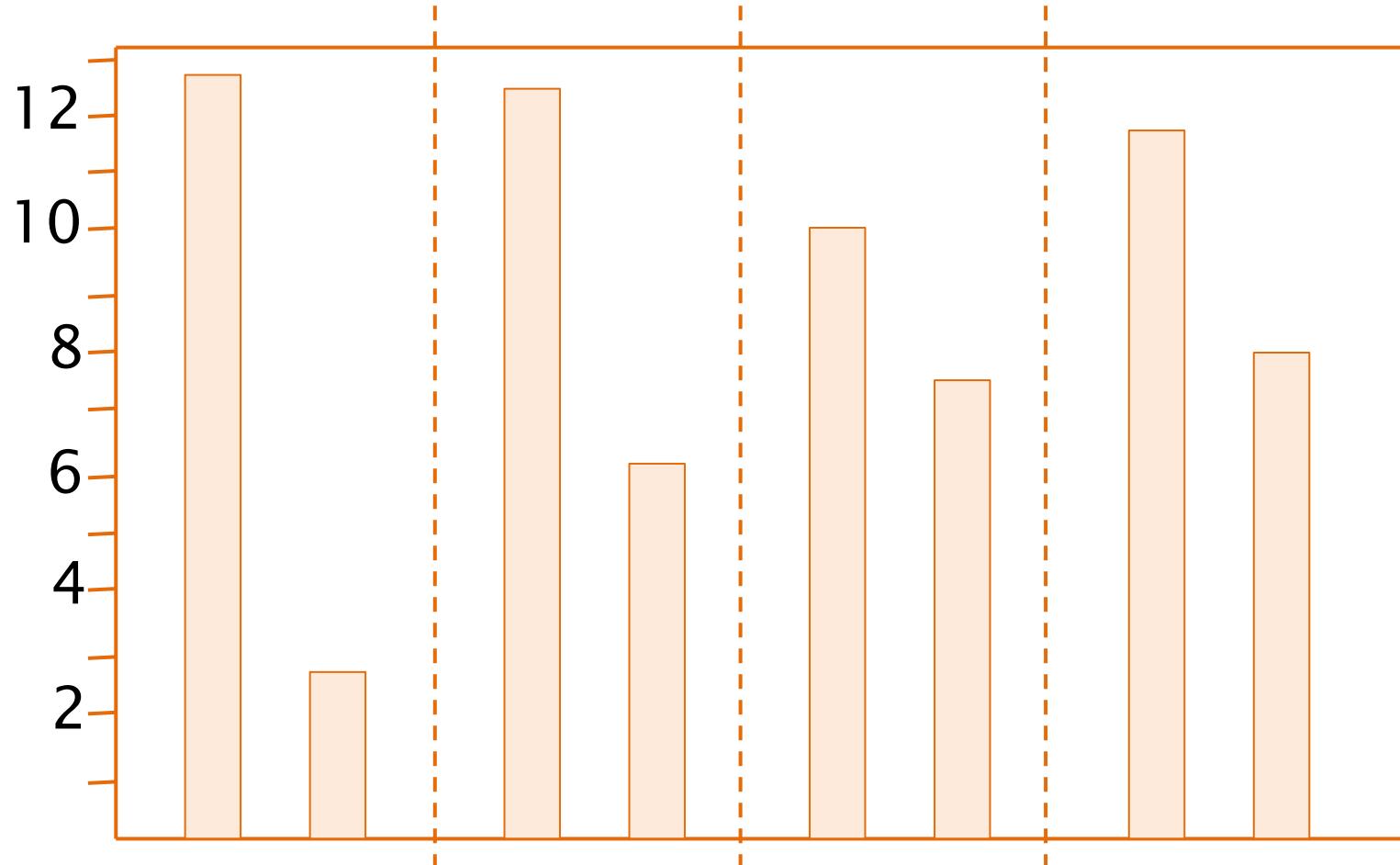
Data-ink ratio = $\frac{\text{data ink}}{\text{total ink used to print the graphic}}$

- Proportion of a graphic's ink devoted to the non-redundant display of data information

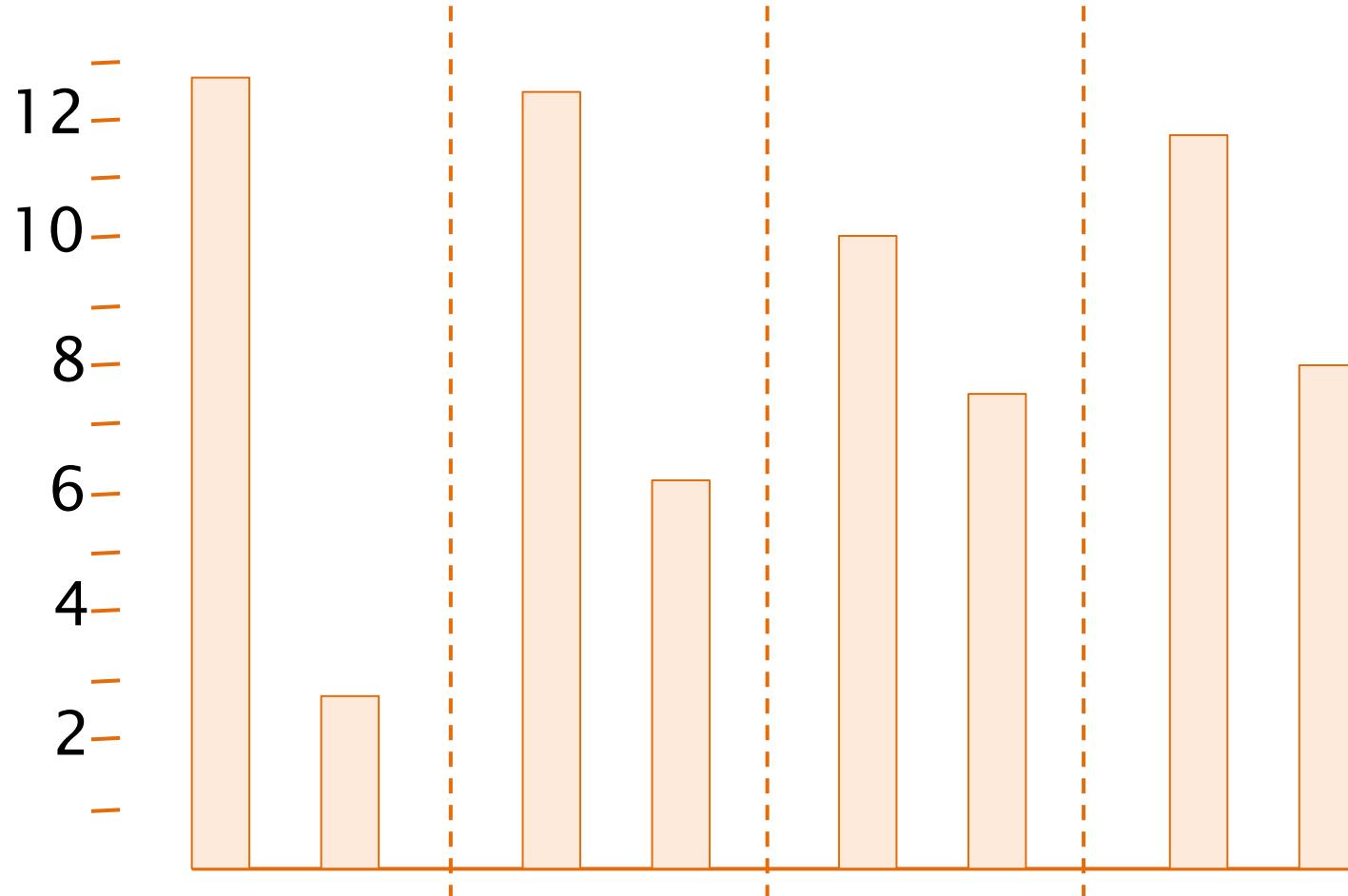
- ◆ Or:

$$1 - \frac{\text{ink that can be erased without loss of information}}{\text{total ink used to print the graphic}}$$

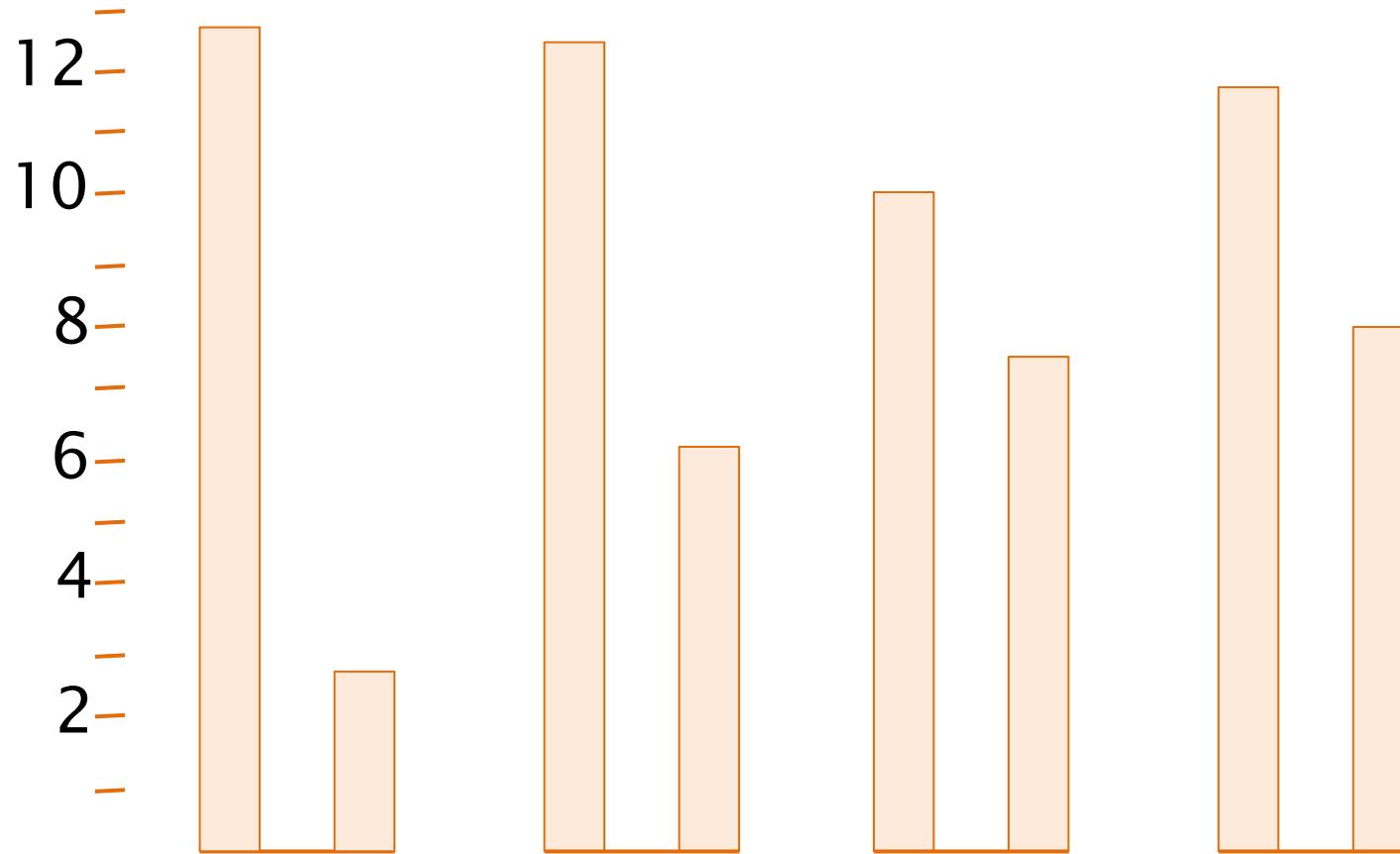
Data-link



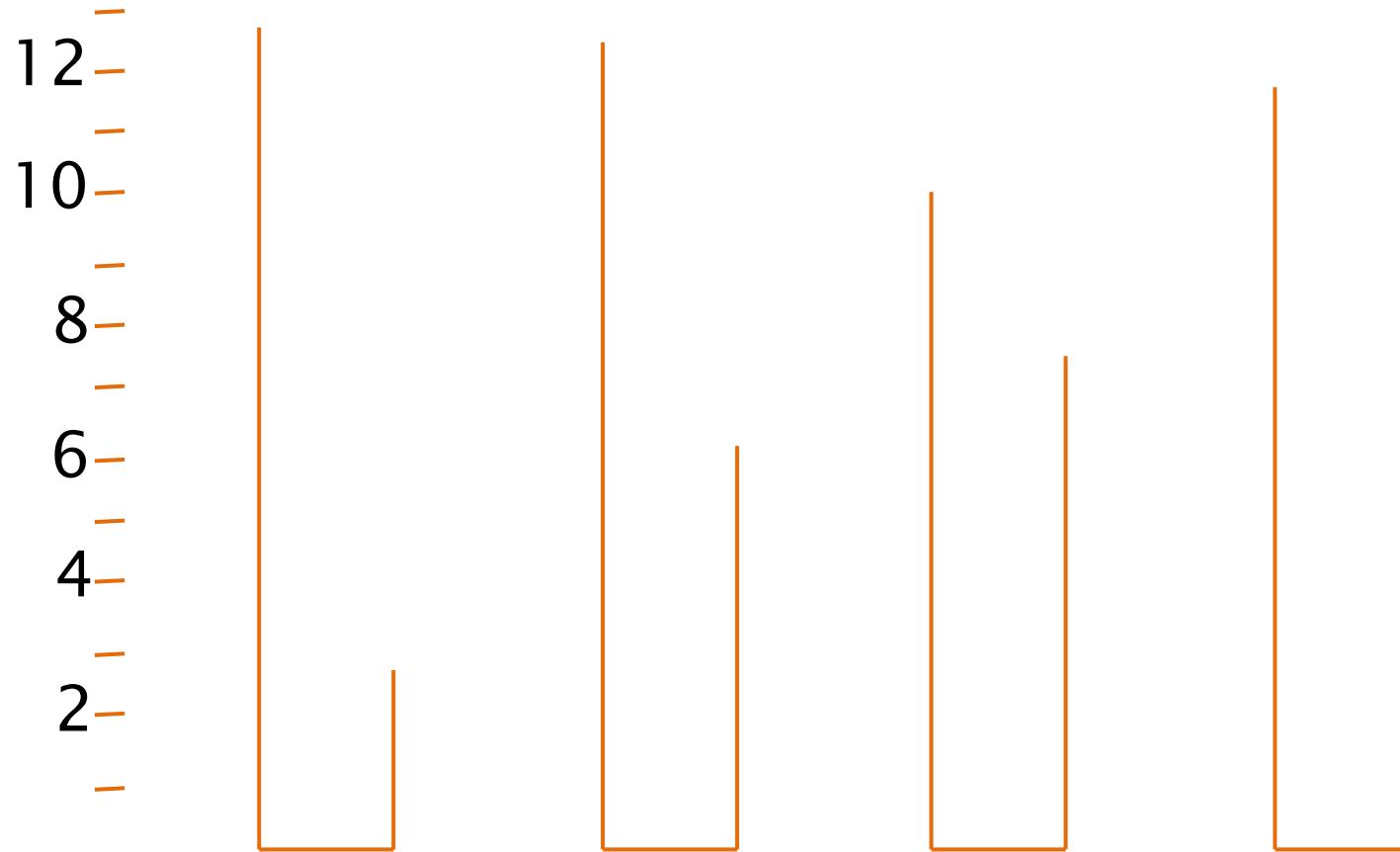
Data-link



Data-ink



Data-ink



Tufte's proposed redesign

Guidelines for design

- Maximize data-ink ratio
 - ◆ Erase non-data-ink
 - ◆ Erase redundant data-ink
- “Within reason”

Above all else show the data

E.Tufte

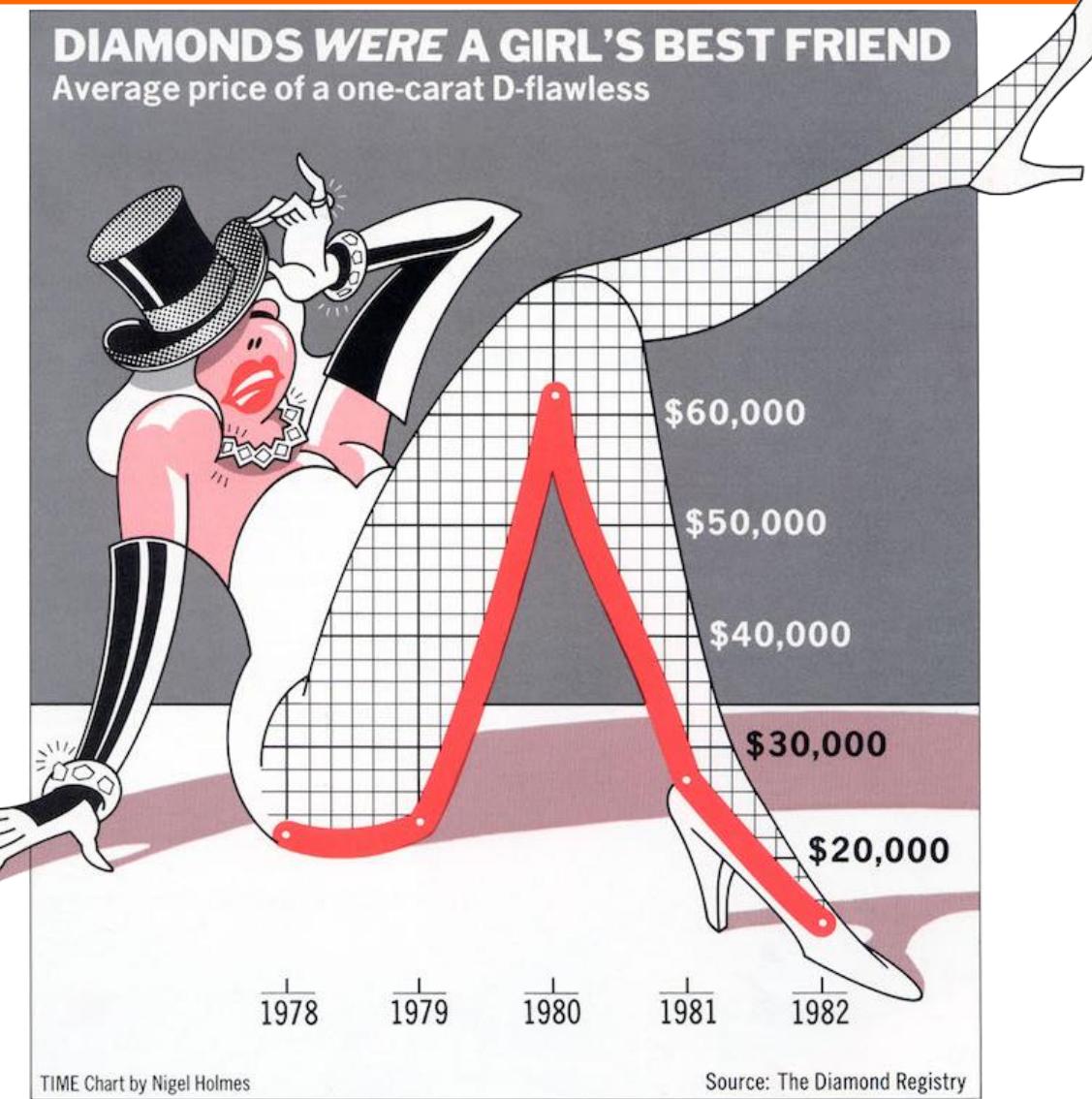
Use of contrast

- Include differences corresponding to actual differences
- Effective when one item is different in a context of other items that are the same
 - ◆ Bright saturated color among mid colors

Chartjunk

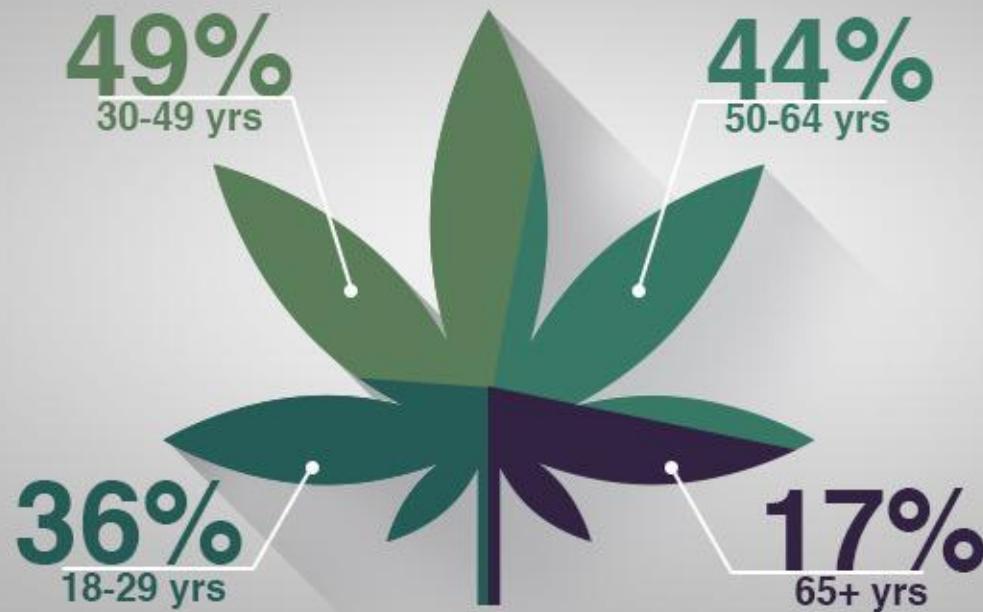
- The presence of unnecessary elements that distract or hide the message conveyed by the diagram

Chartjunk



Nigel Holmes:
<http://nigelholmes.com>

AMERICANS WHO HAVE TRIED WEED



Source: Gallup



SUNDAYS
10^P
ET/PT

#highprofits

Clarity

- Visual encoding and layout should make perception tasks easy and effortless
- Textual and support elements should provide effective support to understanding the information
- Any variation in the graph should represent useful information otherwise it is noise obfuscating the message

Clarity

- **Textual** elements should provide effective support to understanding
 - ◆ Hierarchical
 - Size and position reflects importance
 - ◆ Readable
 - Large enough
 - ◆ Horizontal
 - ◆ Close to data (avoid legends)
- Always label the axes

Colors

- Get it right in black and white
- Use medium hues or pastels
 - ◆ Bright colors distract and tire out
- Use color only when needed to serve a particular communication goal

Cognitive Dissonance

RED

BLUE

GREEN

YELLOW

Detection and Separation

PUC

Efficiency and efficacy of perception tasks is affected by:

- **Detection**

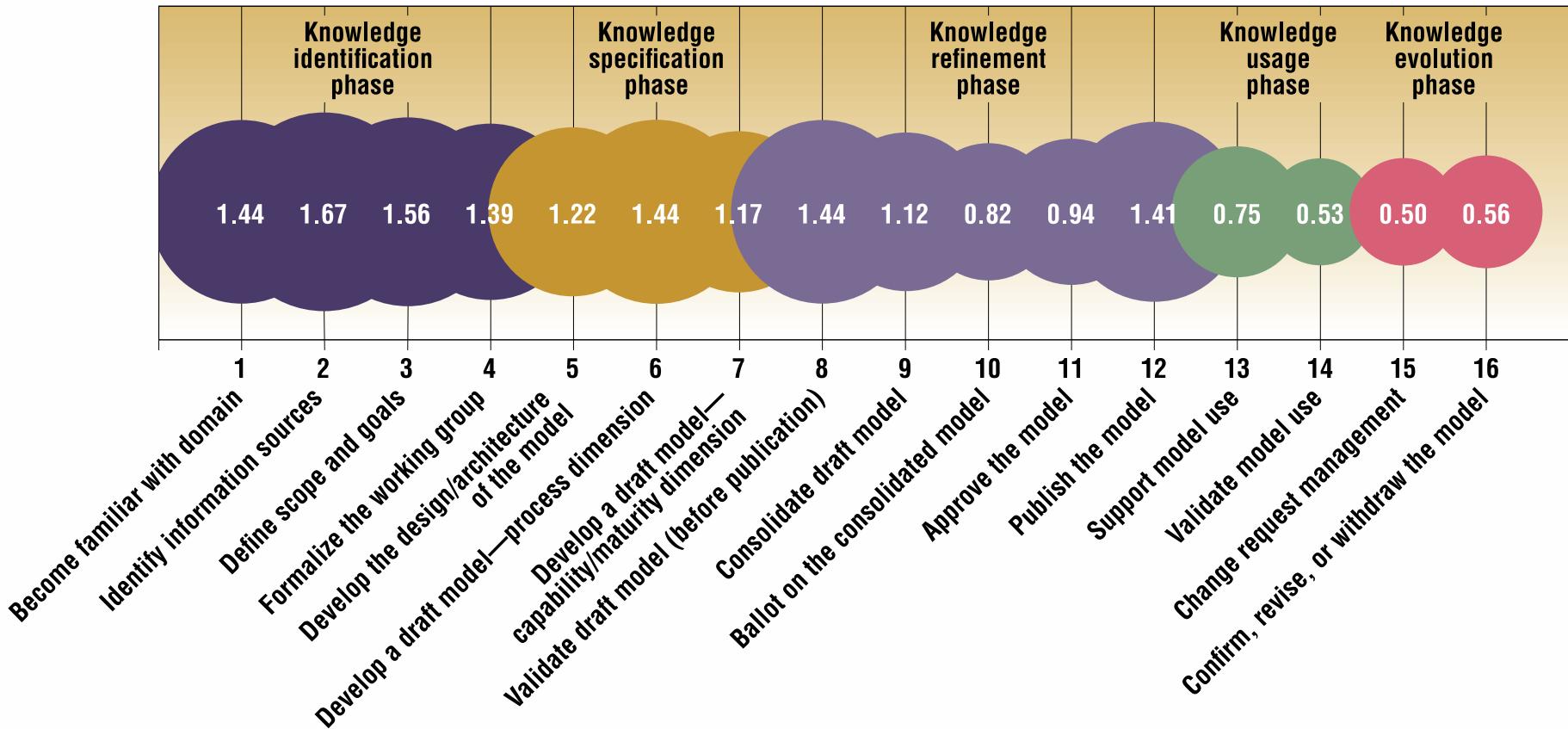
The capability to visually identify the objects that represent the data to be compared

- **Separation**

The distance between the objects to be compared

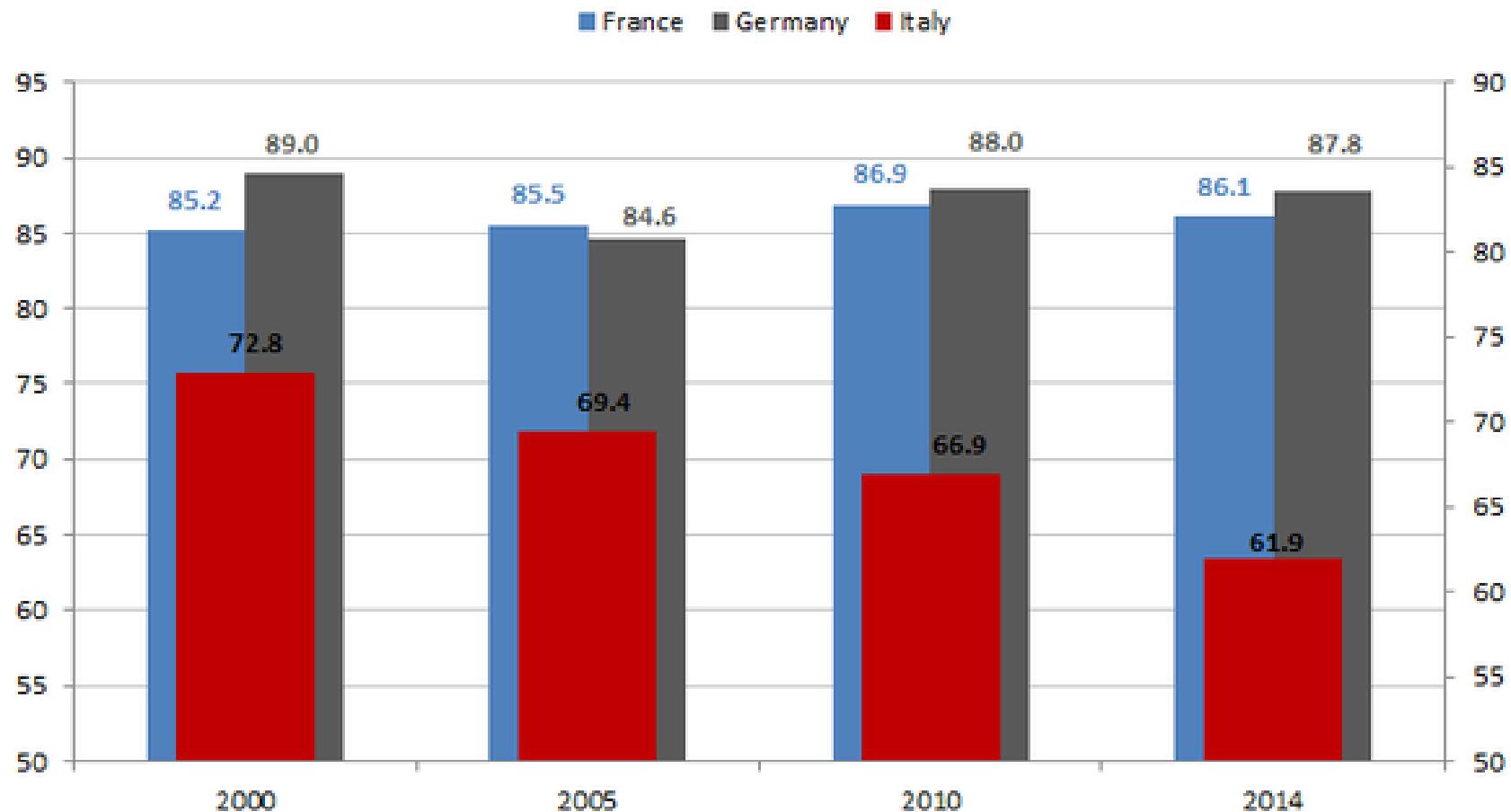
- affects negatively the accuracy

Clarity



Example

Trends in employment rates of 25-34 with a tertiary degree



Analysis

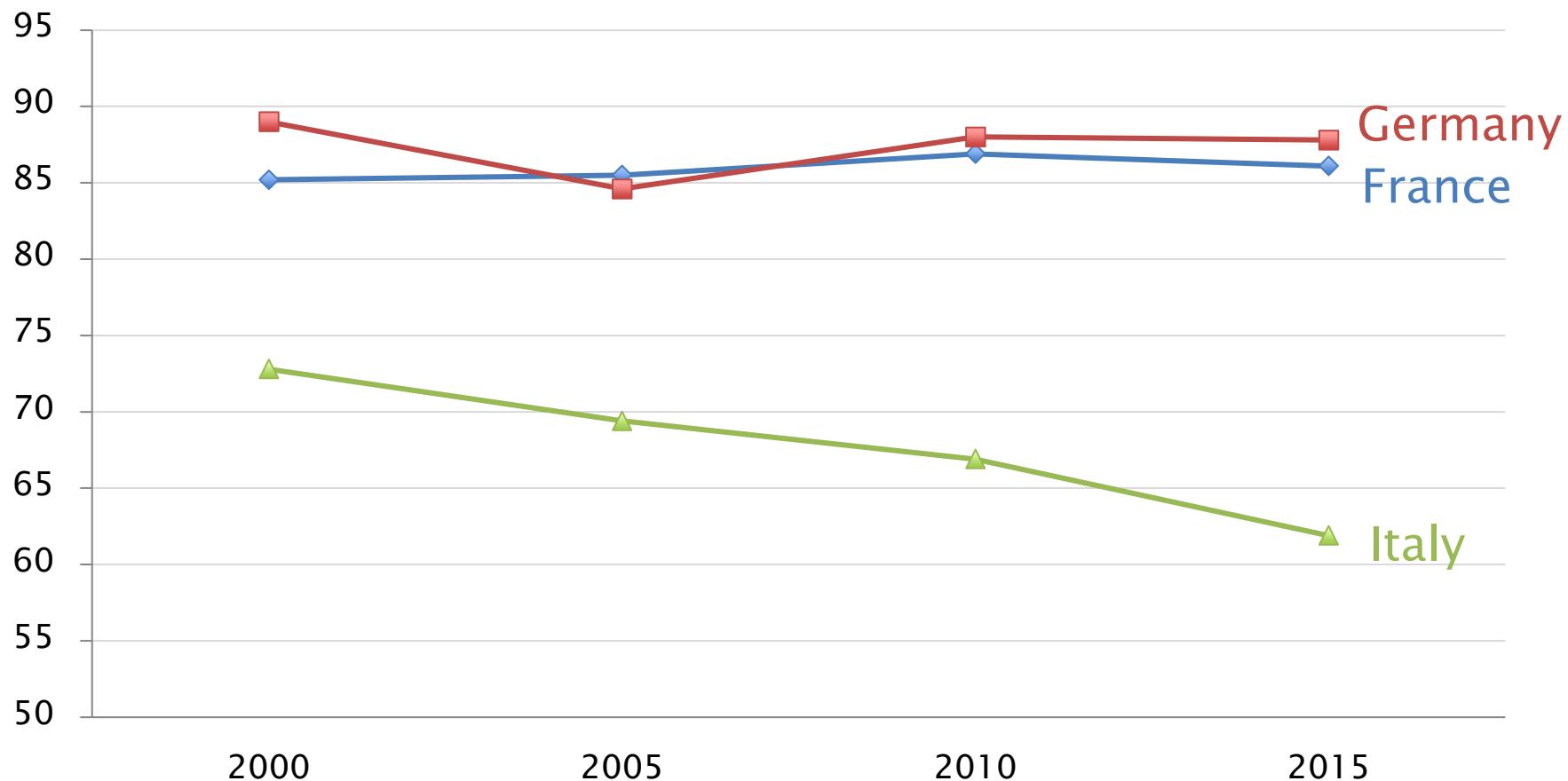
- Proportionality
 - ◆ Due to non-zero base bars, it has a large lie factor (2.2):
 - ratio of real values: 87.8 : 61.9
 - ratio on graph: 37.8 : 11.9
- Utility
 - ◆ Most elements appear useful
 - ◆ X-axis ticks can be removed
 - ◆ Y grid could be made less prominent

Analysis

- Clarity
 - ◆ It uses a **dual scale** that confuses and makes very hard a visual comparison of the values and further distorting the compared values.
 - ◆ The dual scale is not mentioned anywhere and it is not clear which values refer to which scale.
 - ◆ In general the usage of bars is not the most appropriate visual representation if the goal is to show a trend or evolution in time.

Redesign

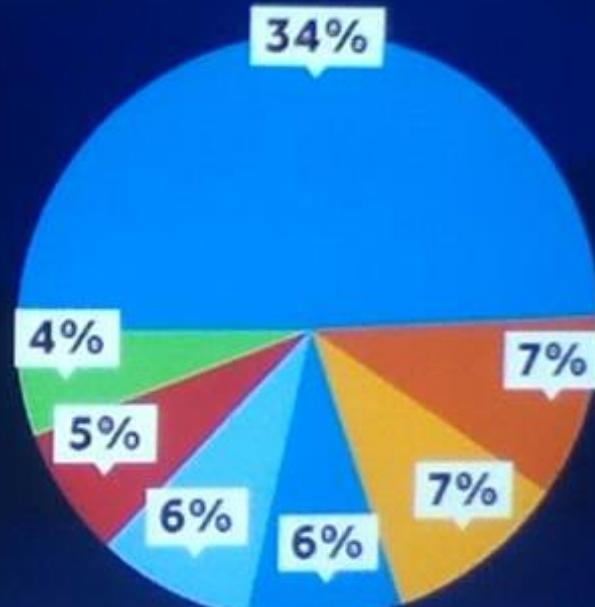
Trends in employment rates of 25–34 with a tertiary degree



Case study

WHICH NFL TEAM IS YOUR FAVORITE?

- PANTHERS
- COWBOYS
- PACKERS
- PATRIOTS
- STEELERS
- REDSKINS
- BRONCOS



SOURCE: PUBLIC
POLICY POLLING

WXII
12

Assessment

- Question:
 - ◆ Is there one (or more) question addressed by the visualization?
- Data:
 - ◆ Is the data quality appropriate?
- Visual Integrity:
 - ◆ Are the visual features appropriate?

Visual Integrity

- Proportionality:
 - ◆ Are the values encoded in a uniformly proportional way?
- Utility:
 - ◆ All the elements in the graph convey useful information?
- Clarity:
 - ◆ Are the data in the graph identifiable and understandable (properly described)?

Question

- What are the most popular/favorite NFL teams in our audience?
- ...

Data

WXII-TV is an NBC-affiliated television station serving North Carolina: home of Panthers

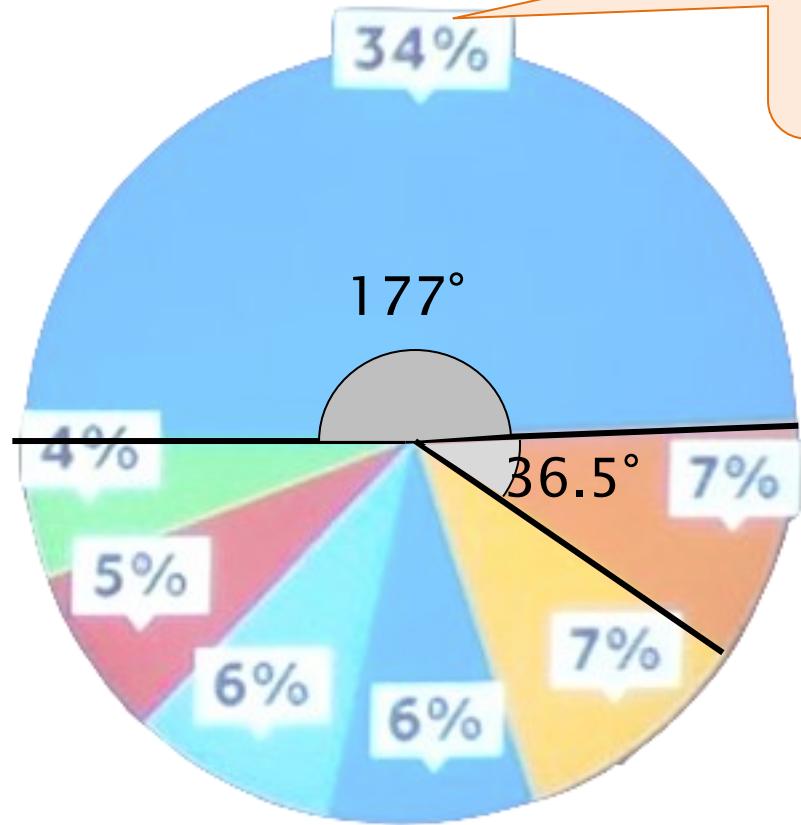
Team	Preferences
Panthers	34%
Cowboys	7%
Packers	7%
Patriots	6%
Steelers	6%
Redskins	5%
Broncos	4%

Total: 69%

Full data

Team	Preferences
Panthers	34%
Cowboys	7%
Packers	7%
Patriots	6%
Steelers	6%
Redskins	5%
Broncos	4%
<i>Other</i>	31%
Total:	100%

Integrity – Proportionality



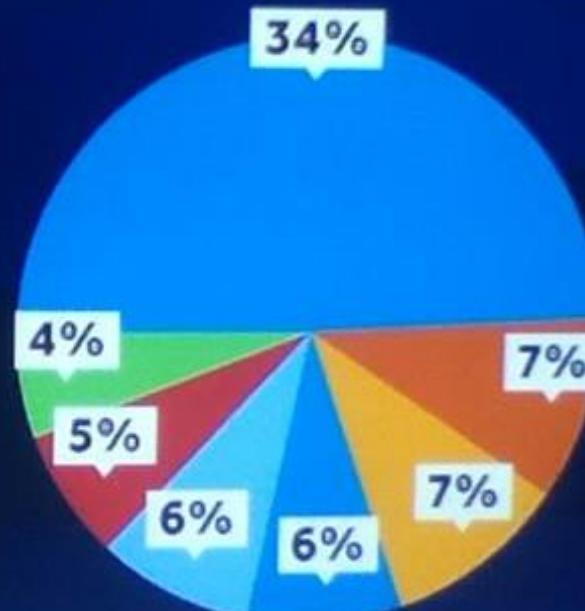
34% corresponds to
50% of the pie!

$$177/36.5 = 4.8$$
$$34/7 = 4.8$$

Utility

WHICH NFL TEAM IS YOUR FAVORITE?

- PANTHERS
- COWBOYS
- PACKERS
- PATRIOTS
- STEELERS
- REDSKINS
- BRONCOS



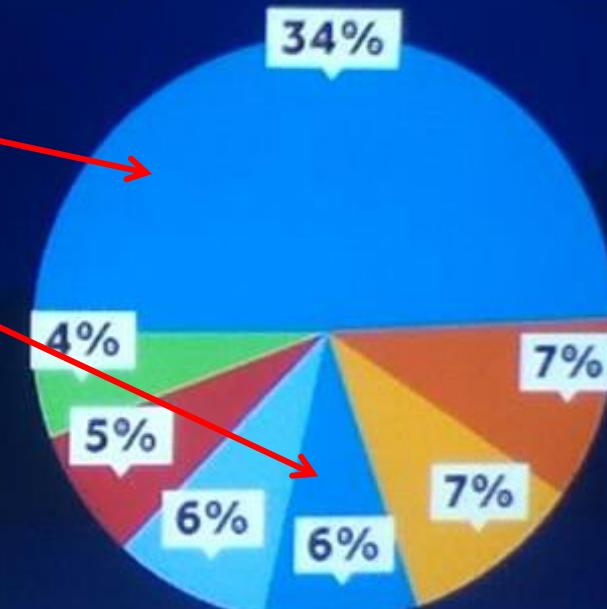
SOURCE: PUBLIC
POLICY POLLING

WXII
12

Clarity

WHICH NFL TEAM IS YOUR FAVORITE?

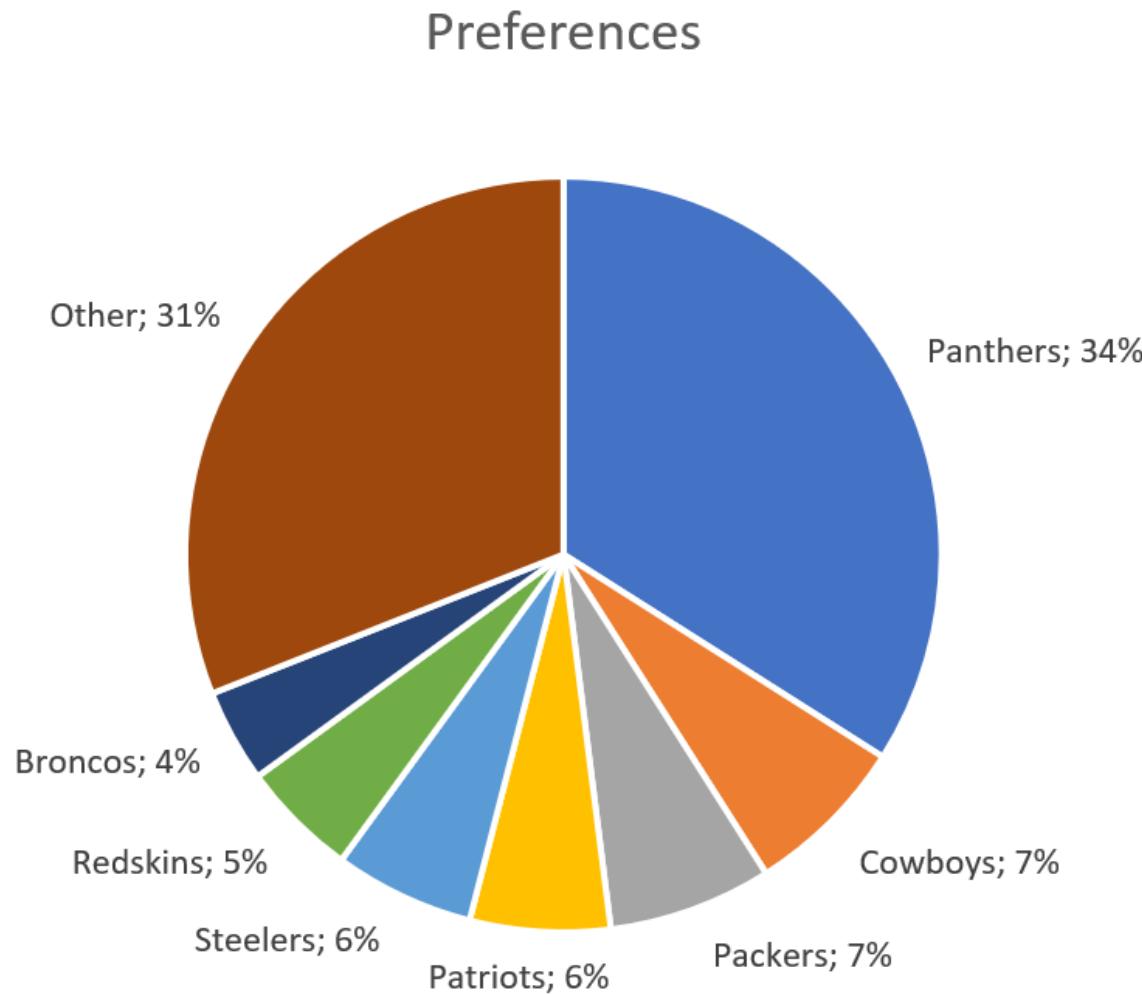
-  PANTHERS
-  COWBOYS
-  PACKERS
-  PATRIOTS
-  STEELERS
-  REDSKINS
-  BRONCOS



SOURCE: PUBLIC
POLICY POLLING

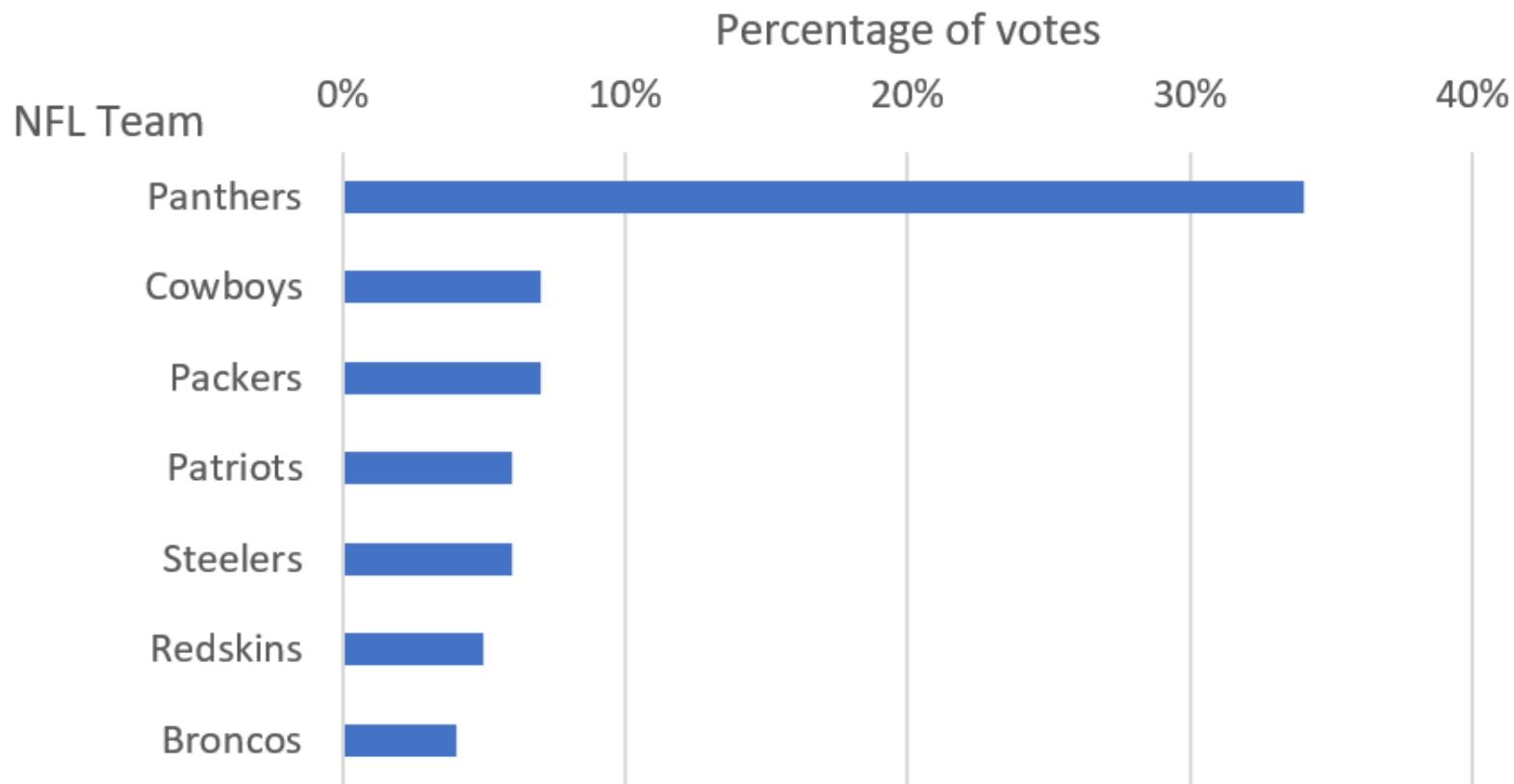
WXII
12

Redesign #1



Redesign #2

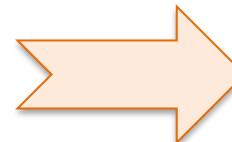
Favorite NFL teams in our audience



VISUALIZATION PIPELINE

Visualization Pipeline

Knowledge



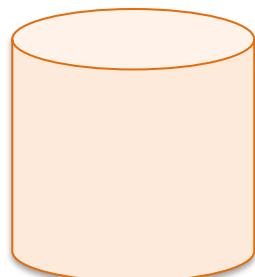
Decisions

Information Understanding
Visual Patterns, Trends, Exceptions

Quantitative Reasoning
Quantitative Relationship & Comparison

Visual Perception
Visual Properties & Objects

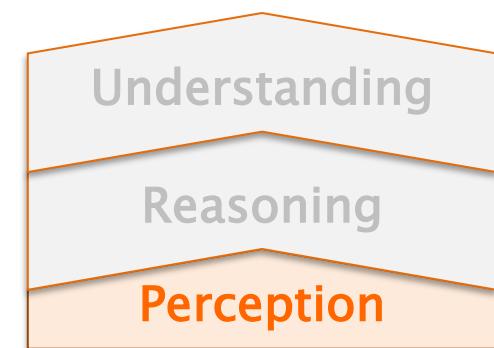
Data



↑ Representation/Encoding

Visual Perception

- Any variable (measure) must be **visually encoded**, i.e. we need to identify:
 - ◆ Visual object to represent entity
 - ◆ Visual attribute to represent the measure



Example

Votes received by four candidates in recent elections

Candidate	Votes	Proportion
Sergio	197800	50.09%
Alberto	140545	35.59%
Giorgio	53748	13.61%
Valter	2759	0.70%

<http://www.comune.torino.it/elezioni/2019/regionali/presidente/citta/>

Encoding

- Visual object: line
- Visual attribute: length

Alberto



- Giorgio

Valter



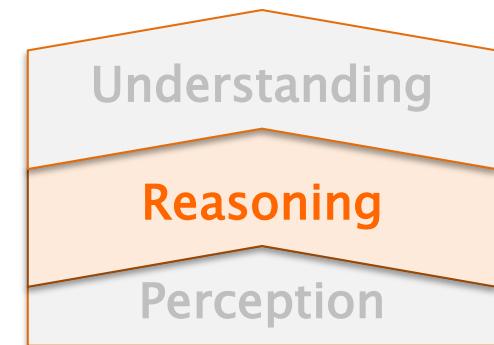
Sergio



Visual Reasoning

Layout and visual attributes allow:

- **Discrimination**
 - ◆ Distinguish visual objects or group of –
- **Comparison**
 - ◆ Place visual objects in order
- **Magnitude assessment**
 - ◆ Evaluate the (relative) magnitude of visual objects



Reasoning

Alberto

- Giorgio
Valter

Sergio

Reasoning

■ Discrimination

Alberto



Valter



Giorgio



Sergio



Reasoning

- Comparison



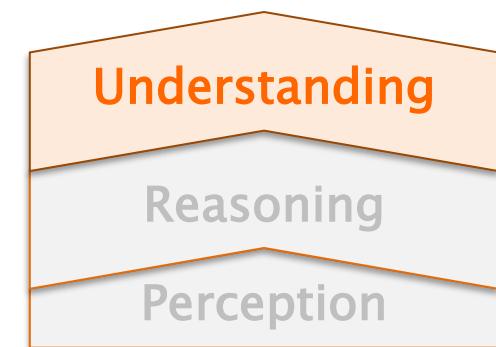
Reasoning

- Assessment



Understanding

- Variation within quantitative measures
 - ◆ Distribution
 - ◆ Deviation
 - ◆ Correlation
- Variation within category
 - ◆ Ranking
 - ◆ Part-to-whole
 - ◆ Time
 - ◆ Space
- Multivariate



Understanding



Understanding

- Ranking



VISUAL PERCEPTION

Data Visualization

Understanding

Information Visualization

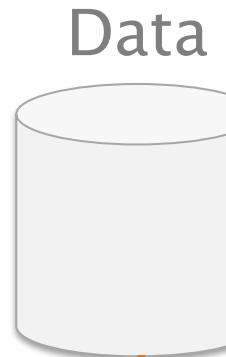
Visual Patterns, Trends, Exceptions

Quantitative Reasoning

Quantitative Relationship & Comparison

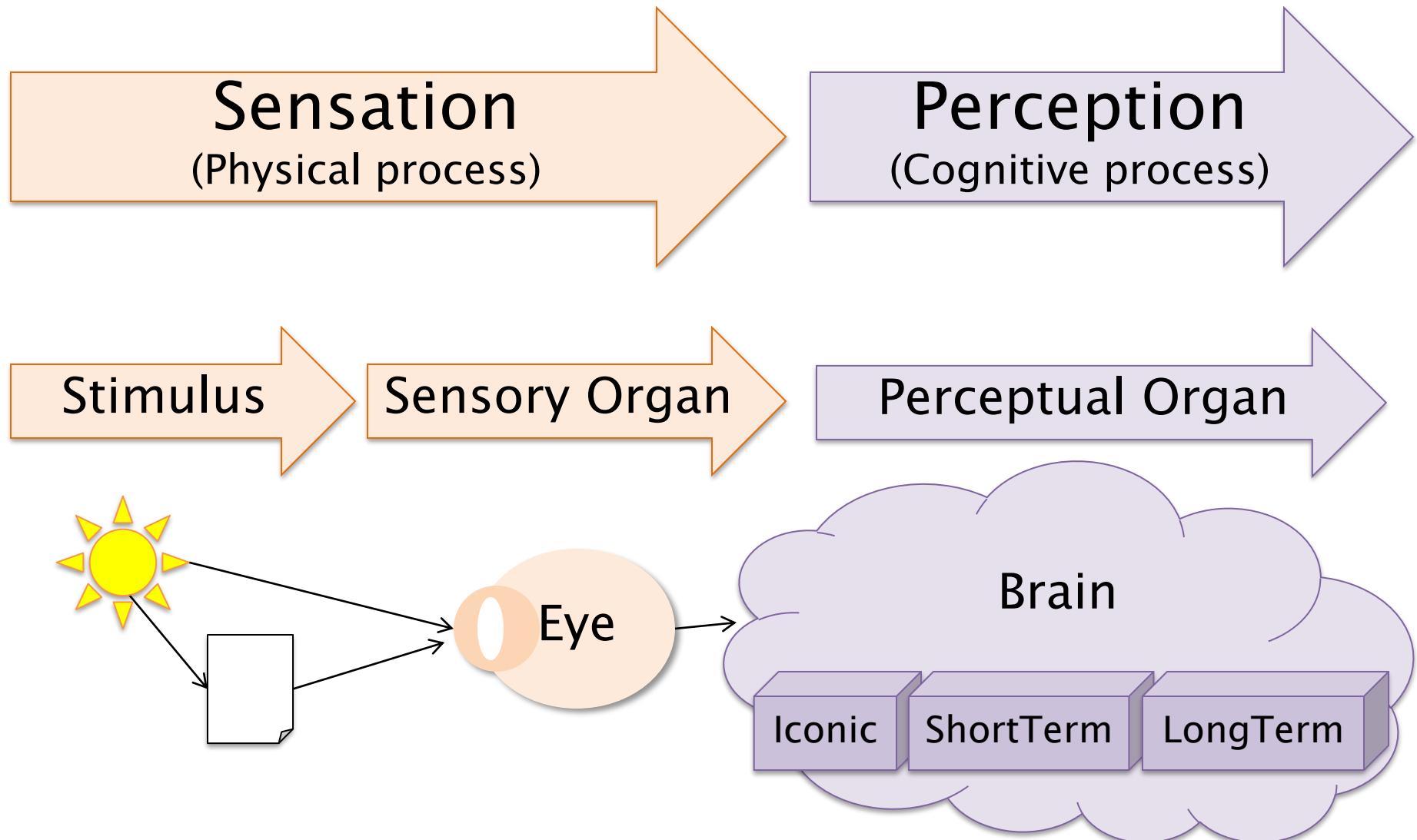
Visual Perception

Visual Properties & Objects



Representation/Encoding

Visual perception



Memory Hierarchy

- Iconic memory (visual sensory register)
 - ◆ Pre-attentive processing
 - ◆ Detects a **limited number of attributes**
- Short-term memory (working memory)
 - ◆ Store visual chunks
 - ◆ Limited number
- Long-term memory
 - ◆ Store high-level knowledge

Simplified Model

- The three levels of memory represent a simplified model
 - ◆ does not correspond to “real” physical brain structure
- Useful to explain a few phenomena
 - ◆ The 7 ± 2 rule
 - ◆ Change blindness

Change blindness



Change blindness



Pre-Attentive Attributes

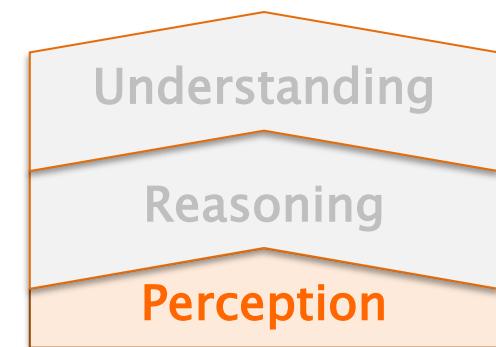
5	7	8	4	9	8	3	1	1	0	6	8	8	2	1	1	5	2	6	6	5
9	5	1	8	4	6	8	4	9	3	0	4	5	3	4	9	2	5	8	5	8
5	0	5	4	6	2	6	5	7	3	7	8	6	5	3	7	2	6	3	1	5
5	8	6	6	8	3	7	6	5	0	9	6	3	4	6	1	9	5	6	6	4
1	6	7	3	9	9	2	8	3	4	0	3	5	1	6	3	5	3	9	3	4
8	6	9	7	5	4	2	4	7	4	9	5	8	5	3	0	7	6	0	6	7
0	3	1	5	3	2	3	5	6	7	2	8	9	8	5	3	7	8	8	2	4
5	5	3	4	8	1	5	6	2	3	5	5	1	2	1	0	8	7	2	6	3
7	4	3	8	4	8	2	6	7	9	5	6	2	3	6	7	8	0	8	3	6
4	9	5	6	7	2	2	2	8	3	1	1	0	1	8	6	2	6	2	1	4

Pre-Attentive Attributes

5	7	8	4	9	8	3	1	1	0	6	8	8	2	1	1	5	2	6	6	5
9	5	1	8	4	6	8	4	9	3	0	4	5	3	4	9	2	5	8	5	8
5	0	5	4	6	2	6	5	7	3	7	8	6	5	3	7	2	6	3	1	5
5	8	6	6	8	3	7	6	5	0	9	6	3	4	6	1	9	5	6	6	4
1	6	7	3	9	9	2	8	3	4	0	3	5	1	6	3	5	3	9	3	4
8	6	9	7	5	4	2	4	7	4	9	5	8	5	3	0	7	6	0	6	7
0	3	1	5	3	2	3	5	6	7	2	8	9	8	5	3	7	8	8	2	4
5	5	3	4	8	1	5	6	2	3	5	5	1	2	1	0	8	7	2	6	3
7	4	3	8	4	8	2	6	7	9	5	6	2	3	6	7	8	0	8	3	6
4	9	5	6	7	2	2	2	8	3	1	1	0	1	8	6	2	6	2	1	4

Encoding

- Encoding is the key to enable visual perception
 - ◆ Visual object to represent entity
 - ◆ Visual attribute to represent the measure
- Two main types
 - ◆ Quantitative (different properties)
 - ◆ Categorical (ordinal or not)



Pre-Attentive attributes

Category	Attribute
Form	Orientation Length/distance Line width Size Shape Curvature Added marks Enclosure
Color	Hue Intensity
Spatial position	2-D position
Motion	Flicker Direction Speed

Perception task

Visual attributes allow:

- Discrimination
 - ◆ Distinguish visual objects
- Comparison
 - ◆ Place visual objects in order
- Magnitude assessment
 - ◆ Evaluate the (relative) magnitude of visual objects

Just noticeable difference

- Given a physical dimension (length, brightness, etc.) x
- d is the **just noticeable difference** if:
 - ◆ difference between x and $x+d$ is perceivable
 - ◆ but not smaller differences
- d depends on many factors:
 - ◆ Subject
 - ◆ Environment
 - ◆ Physical dimension

Weber's law

- Just noticeable difference d is:

$$d_p(x) = k_p \cdot x$$

- Where
 - ◆ x : dimension
 - ◆ $d_p(x)$: just noticeable difference
 - ◆ k_p : constant
 - Subjective
 - Environmental

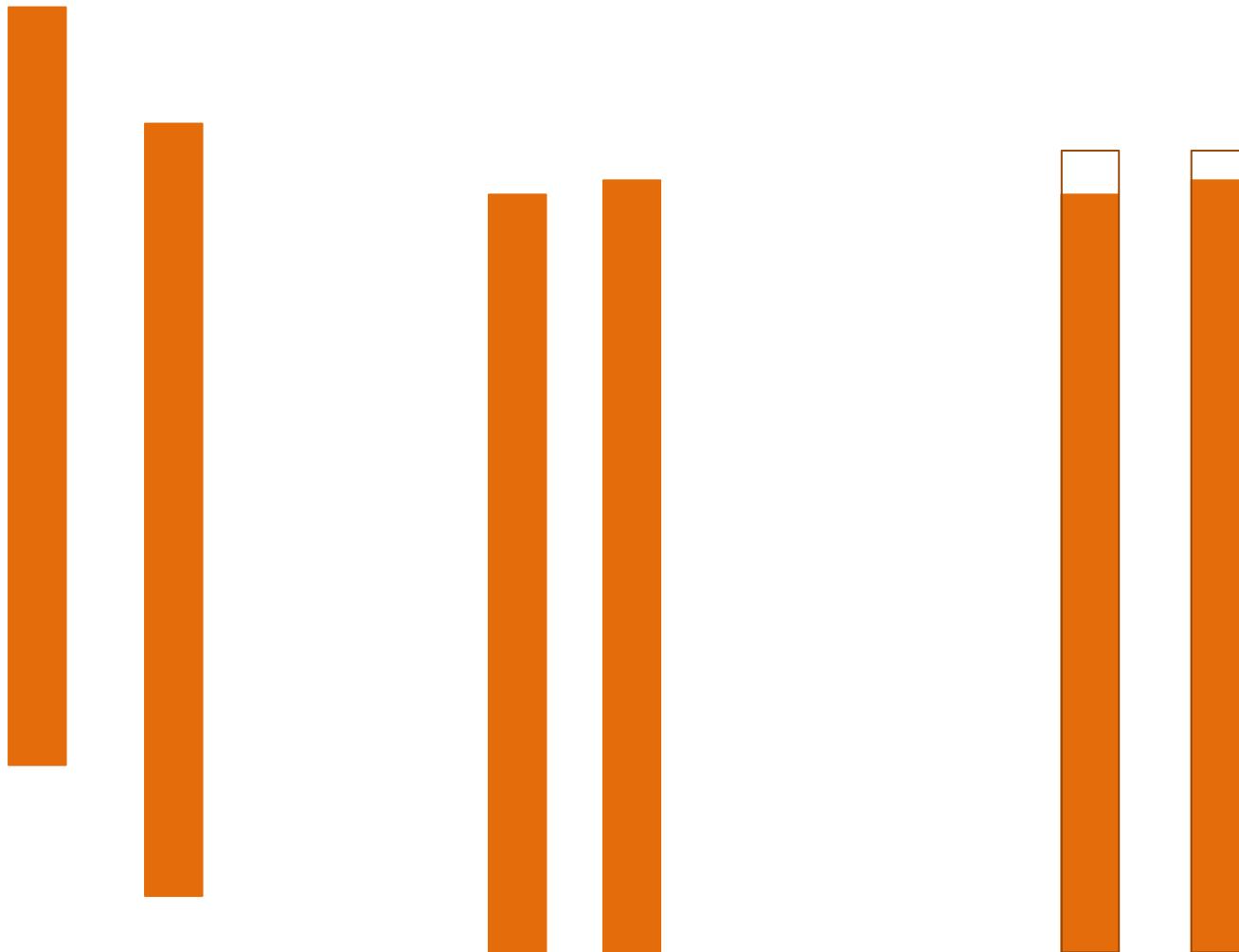
Consequences of Weber's law

- It is easier to compare lengths that differ by a large percentage
- The same difference is easier to notice between smaller measures
 - ◆ More likely to be larger than just noticeable difference

$$x < y \implies d_p(x) < d_p(y)$$

- Length of non-aligned objects is harder to compare
 - ◆ Double comparison

Non-aligned objects lengths



Non-aligned objects lengths

- Additional references may help comparison
 - ◆ They provide alternative possible comparisons
- If lengths range between 0 and a maximum (L), e.g. percentages
- Comparing l_1 and l_2 (close to L) that differ by a small amount d
 - ◆ Difference $L-l_1$ vs. $L-l_2$ easier to notice than l_1 vs. l_2

Stevens's law

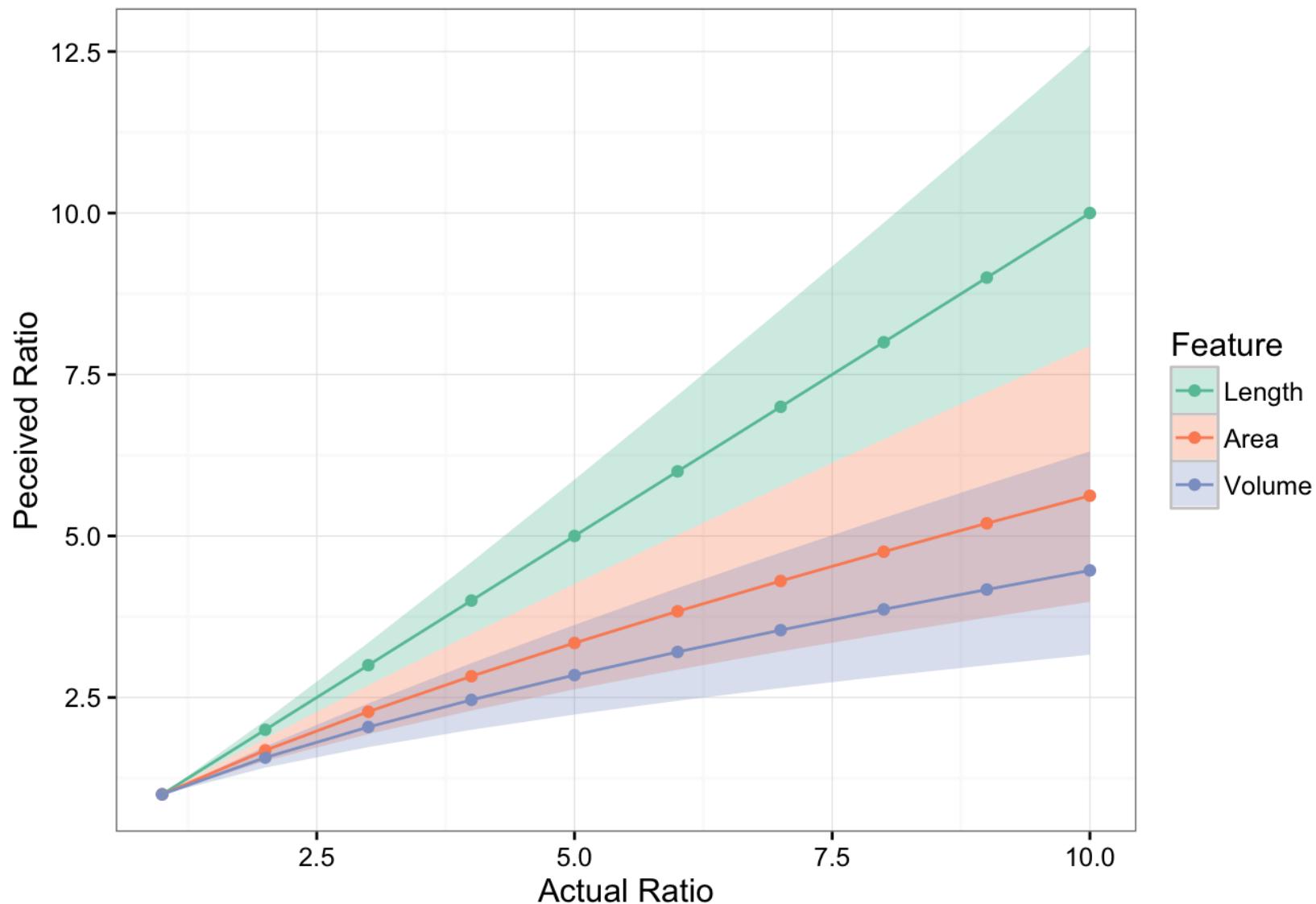
- Perceive scale (magnitude ratio)

$$p(x) = c \cdot x^\beta$$

- Where β depends on spatial dimension
 - ◆ 1D: Length $\rightarrow \beta$ in [0.9, 1.1]
 - ◆ 2D: Area $\rightarrow \beta$ in [0.6, 0.9]
 - ◆ 3D: Volume $\rightarrow \beta$ in [0.5, 0.8]

Stevens S. S. (1975). Psychophysics, John Wiley & Sons.

Stevens's law



Stevens's law

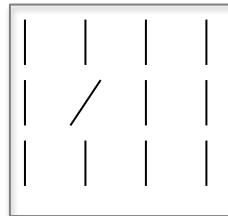


Consequences

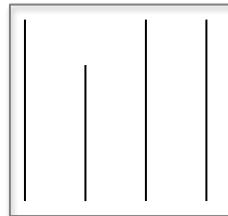
- Prefer comparing lengths
- Avoid comparison between areas
 - ◆ Except for ordinal measures
- Never-ever make volume comparisons

Attributes of form

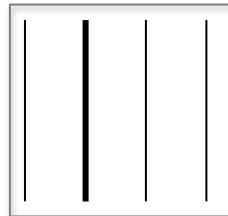
Orientation



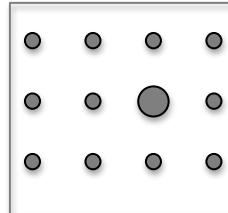
Line Length



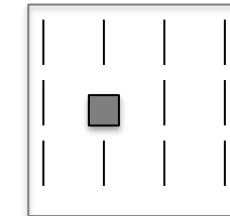
Line Width



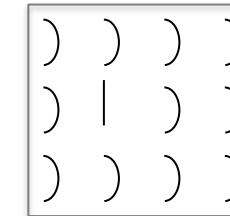
Size



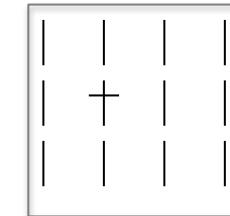
Shape



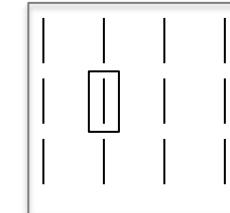
Curvature



Added mark

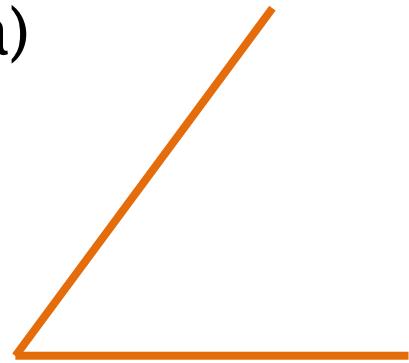


Enclosure

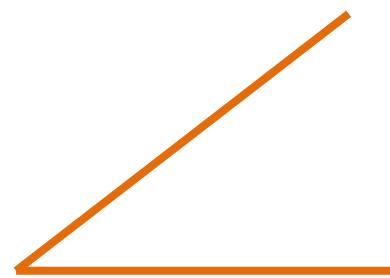


Orientation (angle or slope)

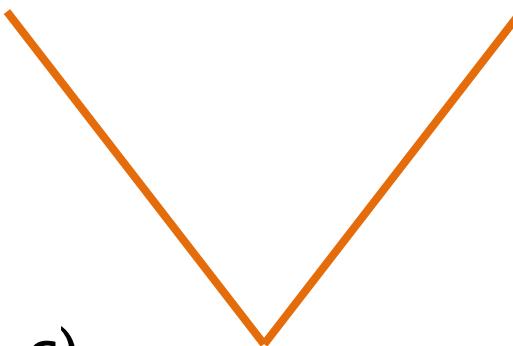
a)



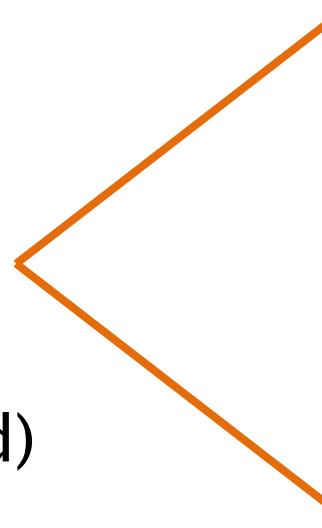
b)



c)



d)

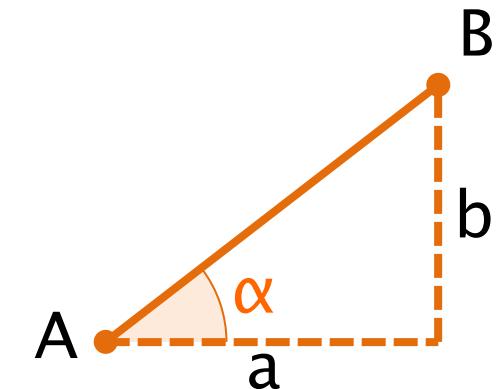


Angle vs. Slope

- Slope of A–B is b/a

- ◆ $\tan(\alpha)$

- Slope judgment typically falls back to an angle judgment
 - ◆ Given an error ϵ in the angle judgment
 - ◆ It is reflected in a slope error

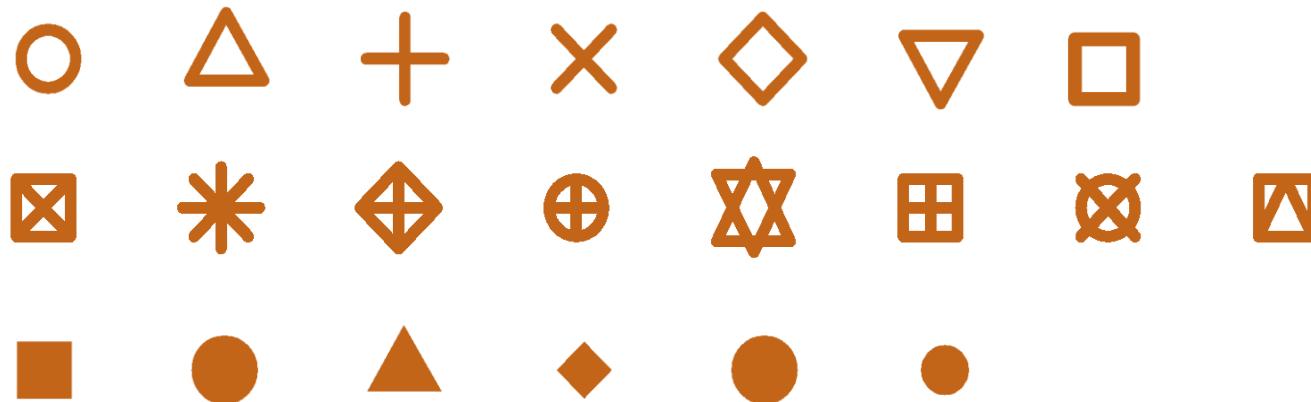


$$\tan(\alpha + \epsilon) - \tan(\alpha) = \epsilon \cdot \tan'(\alpha) = \frac{\epsilon}{\cos^2(\alpha)}$$

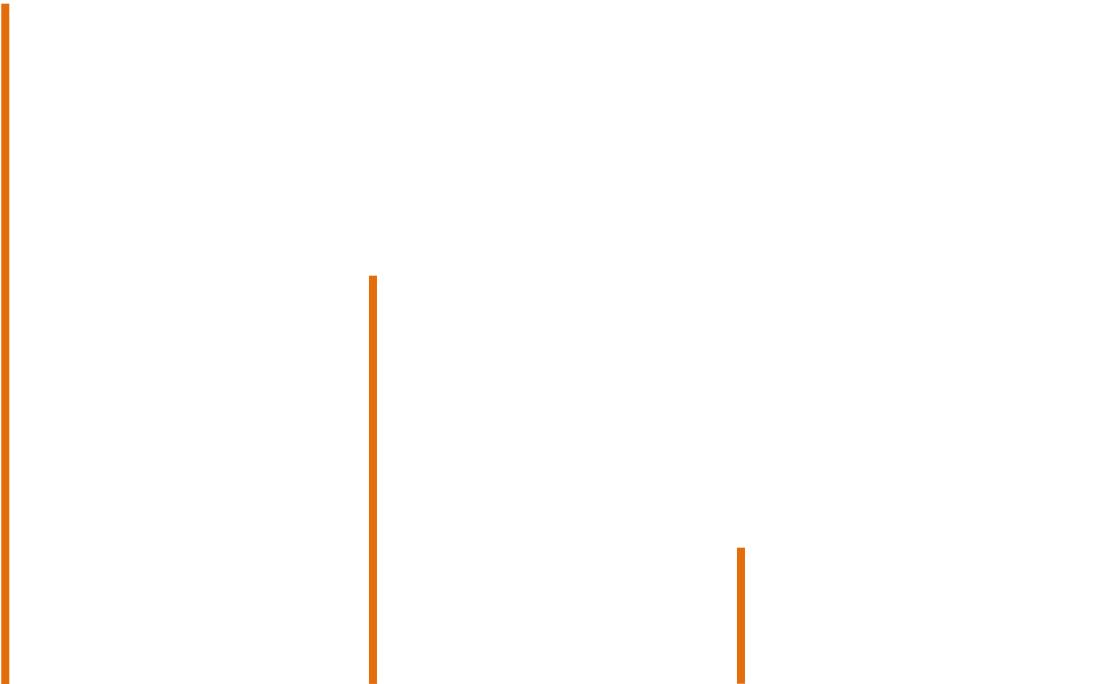
– Getting infinite as α approaches to $\pi/2$

Shape

- There is no common quantitative semantics for the shapes
 - Unless they are characters...
 - ◆ Fill textures are shapes too



Length

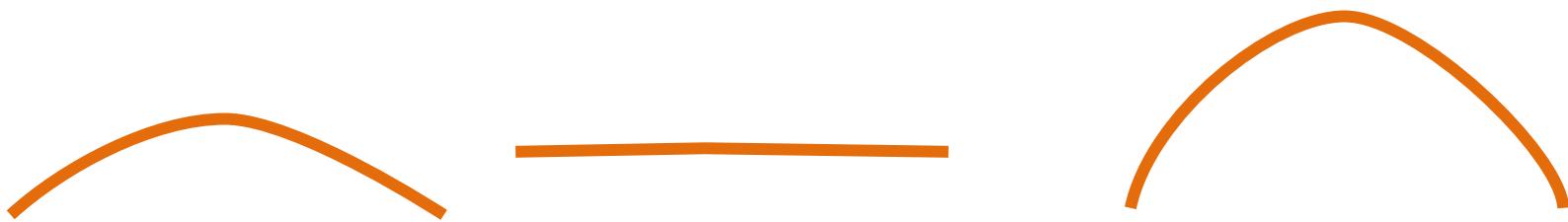


Effect of context



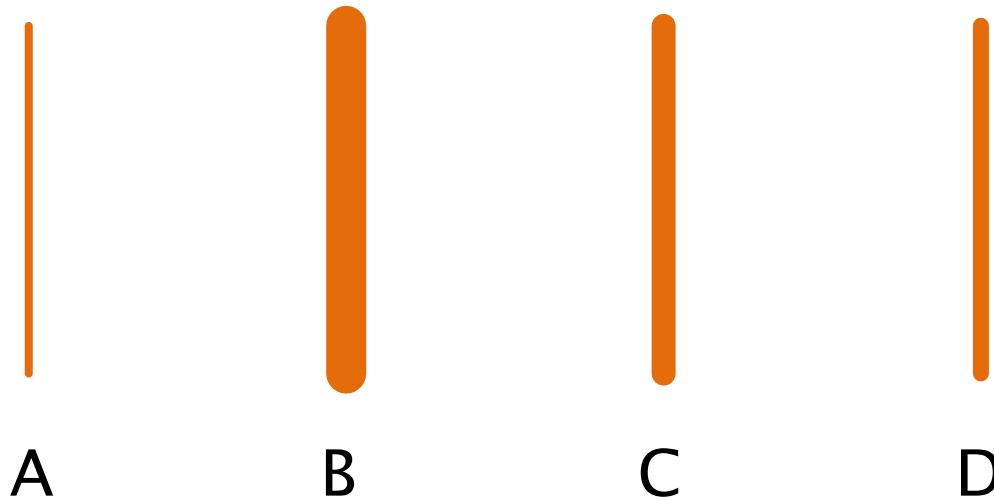
Curvature

- There is no common magnitude assessment for the curvature



Width

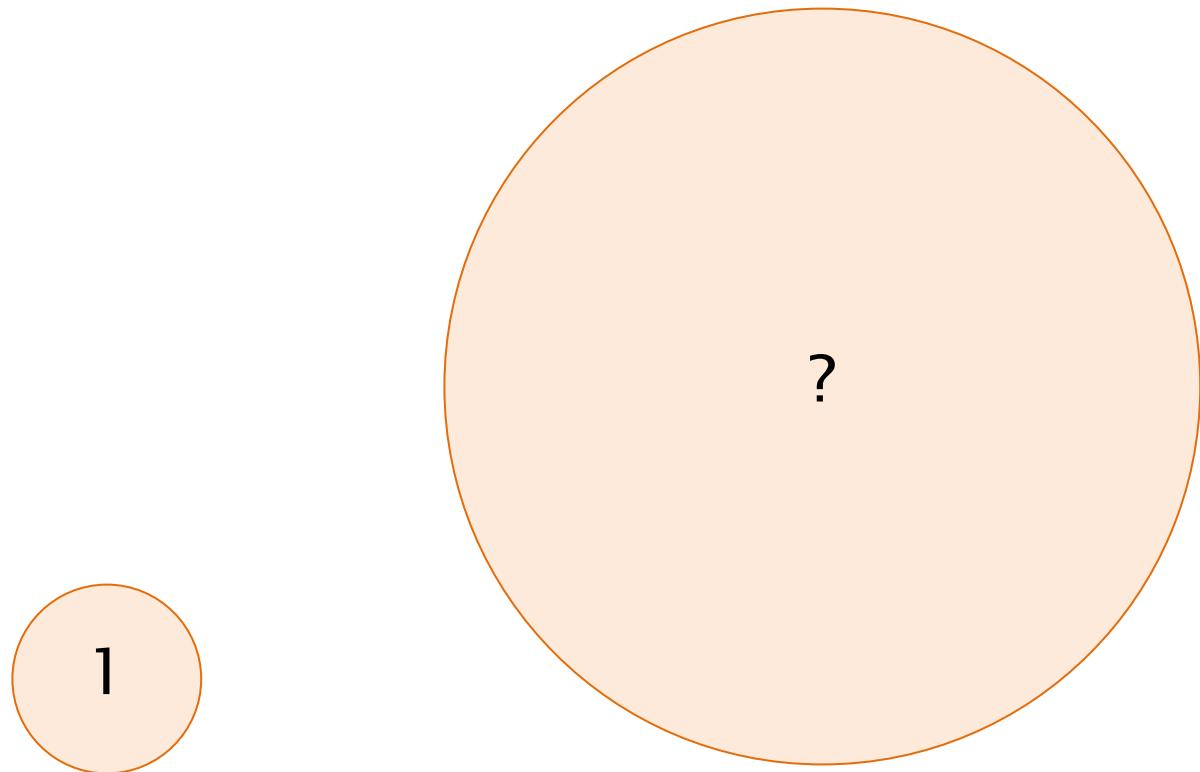
- Order can be identified
 - ◆ Difficult to appreciate actual magnitude



Mark

- No common quantitative semantics of marks
- Number of marks could encode a natural number
 - ◆ Harder to read than a cipher

Size / Area

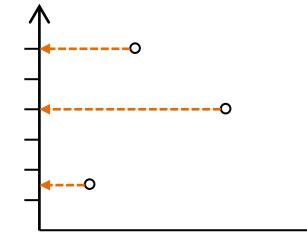


Enclosure

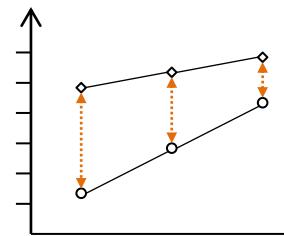
- No common quantitative semantics for enclosure
 - ◆ Except counting items enclosed

Spatial Position

- Position along axis
 - ◆ Common scale
 - ◆ Distinct identical scales
 - Possibly un-aligned

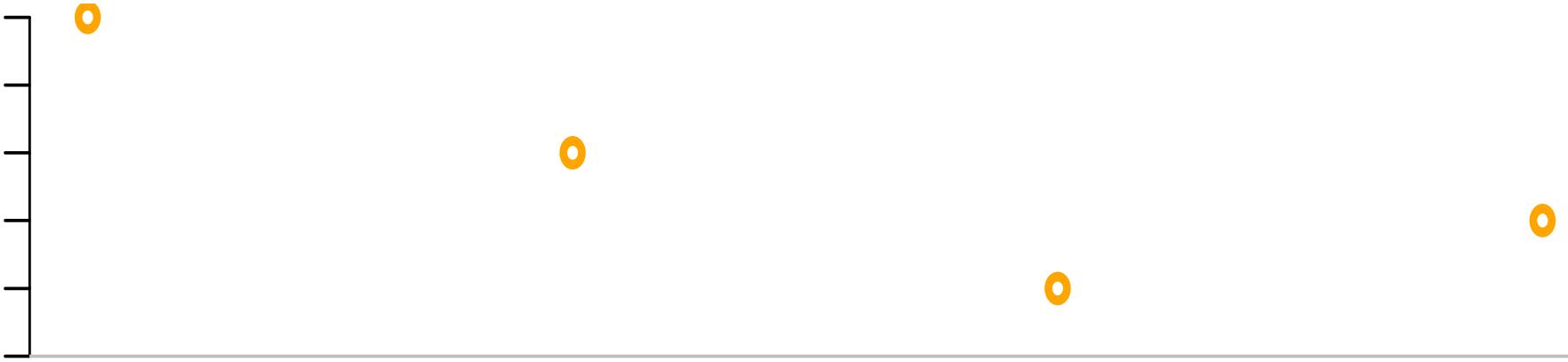


- Distance

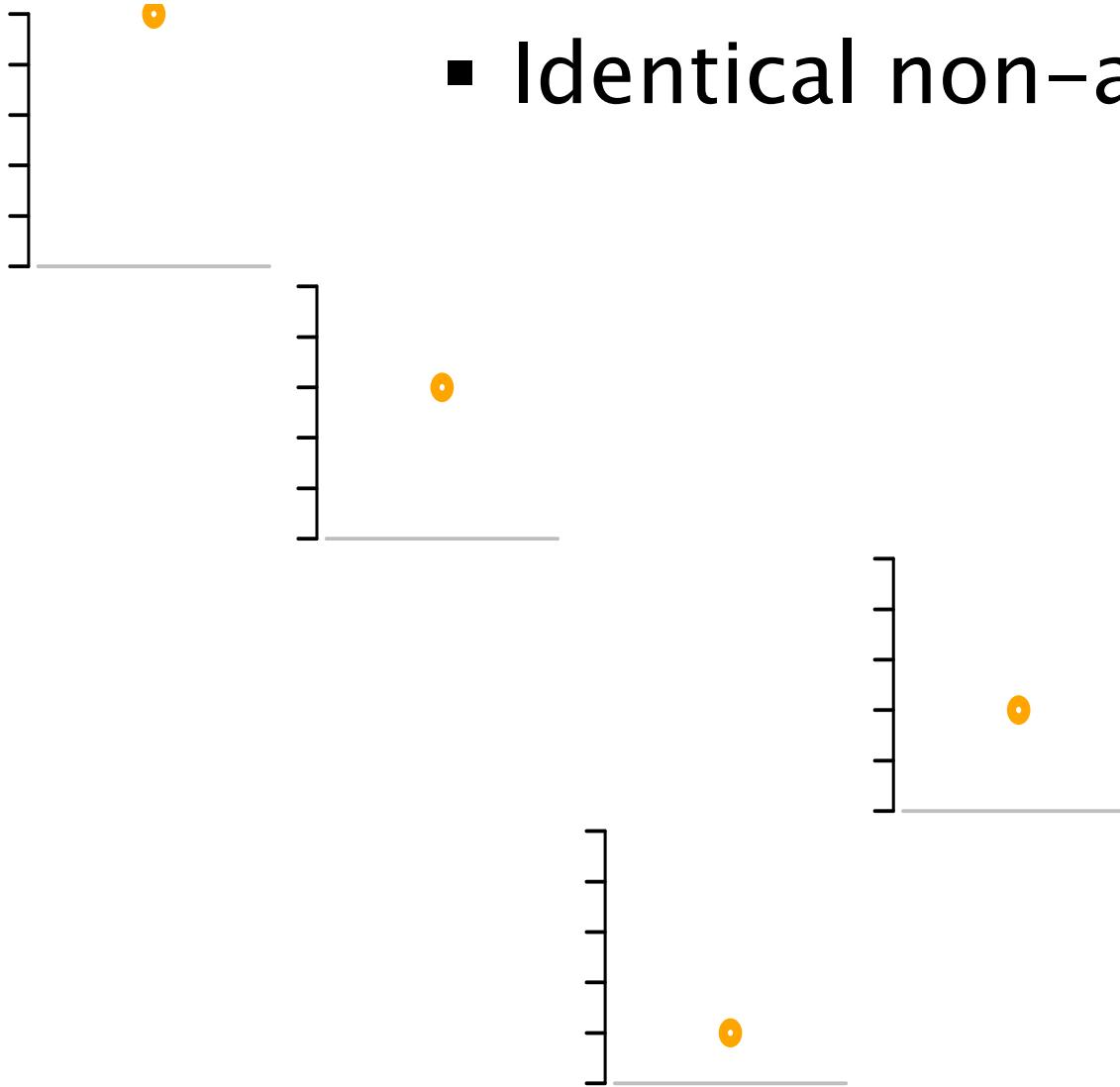


Position

- A common scale



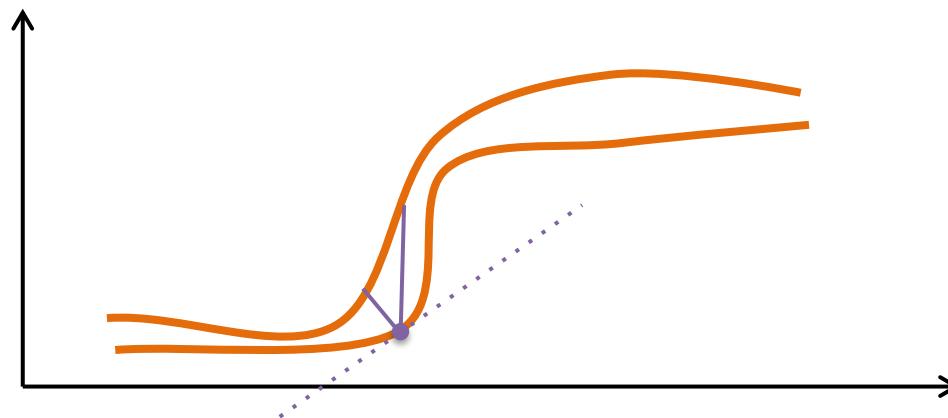
Position



- Identical non-aligned scales

Distance

- Points
 - ◆ Use length of imaginary connecting lines
- Lines
 - ◆ Distance orthogonal to tangent
 - Not what is meant in xy plots



Detection and Separation

Comparison is affected by:

- Detection
 - ◆ The capability to visually identify the objects that represent the data to be compared
- Separation
 - ◆ The distance between the objects to be compared
 - affects negatively the accuracy

Attributes of color

- Hue



- Saturation



- Intensity



- ◆ Luminance
- ◆ Value

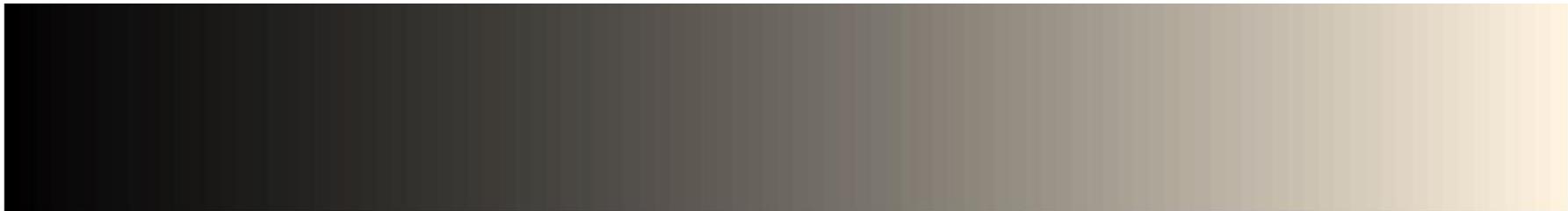
Hue

- There is no common ordering semantics for hues
 - ◆ High spatial frequencies are perceived through intensity changes
 - ◆ Often perceived as separated into bands of almost constant hue, with sharp transitions between hues
- Nominal values can be represented by suitably spaced values



Intensity

- ◆ a.k.a. Luminance, Value
- Provides a perceptually unambiguous ordering
 - ◆ Context can affect accuracy

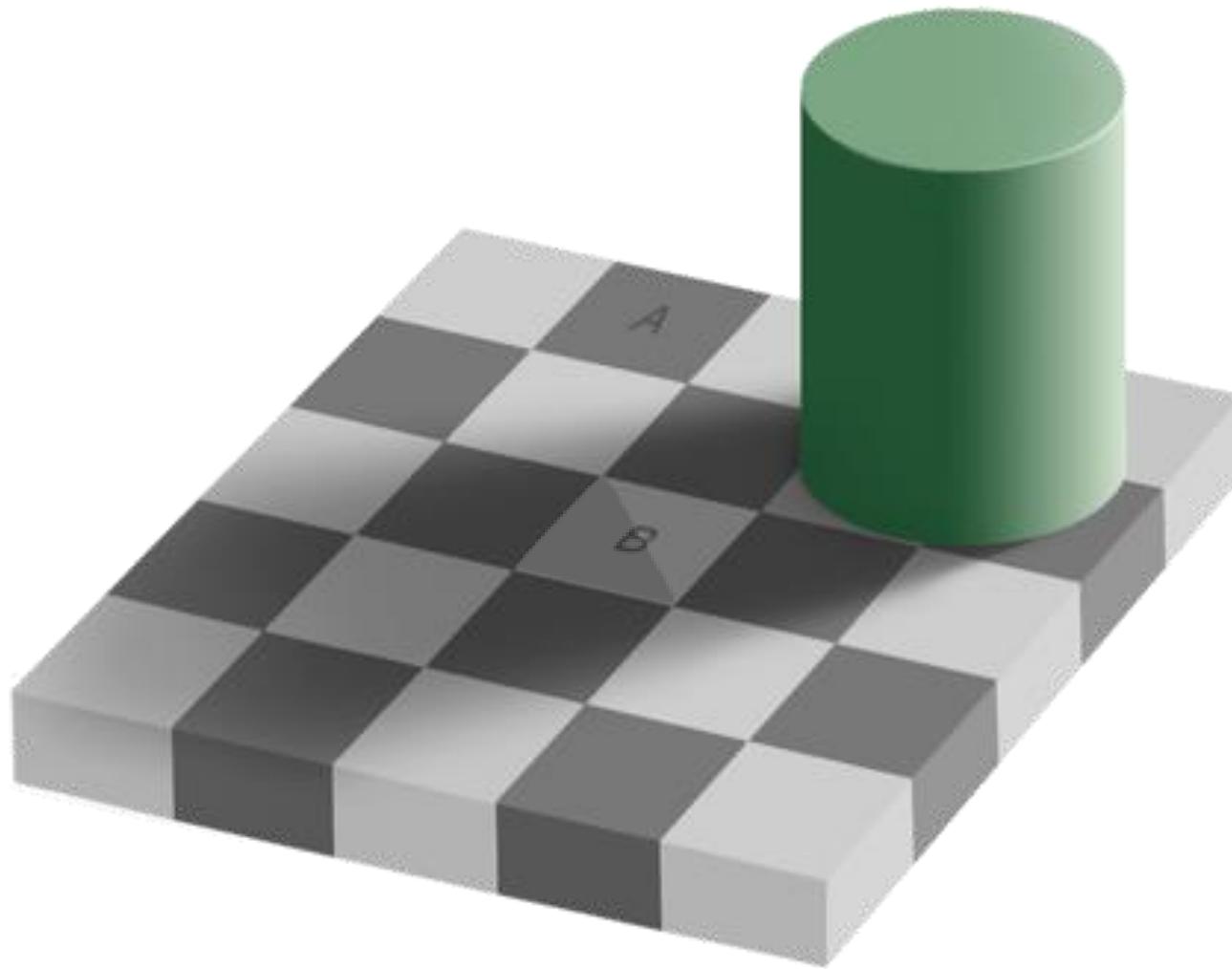


Saturation

- Perceptually difficult to associate an ordered semantics
 - ◆ Can be combined with hue to increase discrimination



Effect of Context



Effect of Context

- Use uniform background
 - ◆ To make distinct visual objects for the same feature look the same
- Use a background color that is contrasting enough with the visual objects' color
 - ◆ To make visual objects easily seen
- Avoid non-uniform background

Color usage

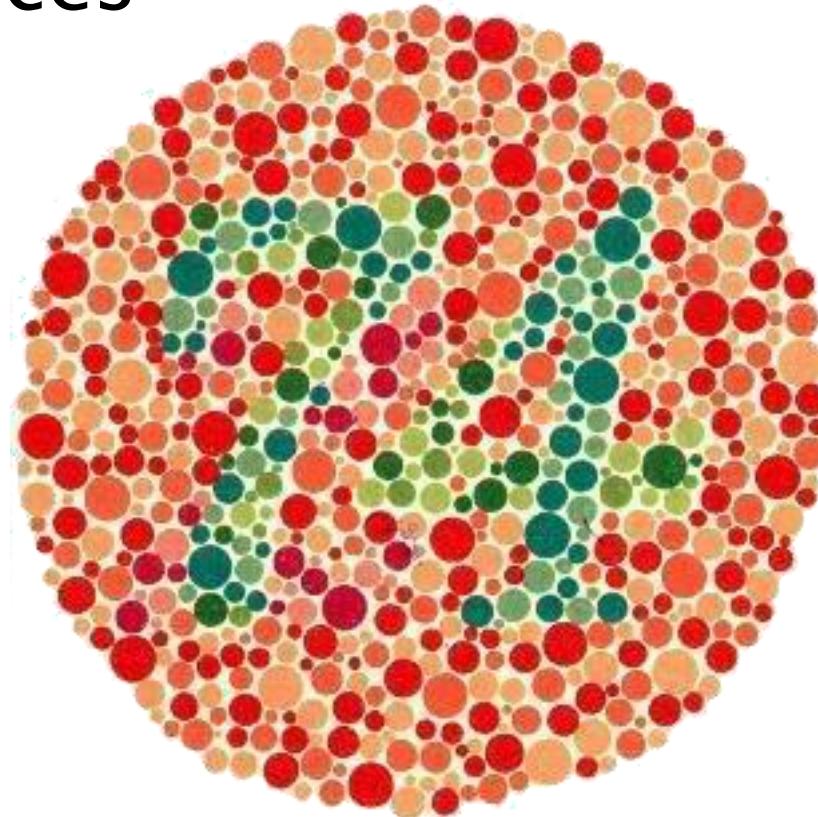
- Ordinal measure should be mapped to increasing saturation **and** intensity
 - ◆ Avoid rainbow palette
- Use sequential or diverging palette
 - ◆ E.g.



- <http://colorbrewer2.org/>

Color Blindness

- Inability to see colors or perceive color differences



<http://www.color-blindness.com>

Visual Encoding: Quantitative

Object	Attribute
Point	Position (w.r.t. axis/axes)
Line	Length Position (w.r.t. axis/axes) Slope
Bar	Length
Shape	Size (area) Count

Visual Encoding: Categorical

Attribute

Position

Size

Color

Shape

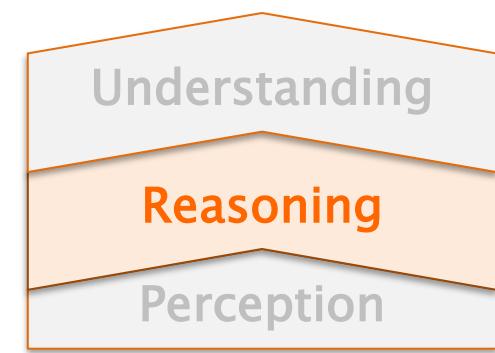
Fill pattern

Line style

Intensity
Saturation
Hue

Ordinal

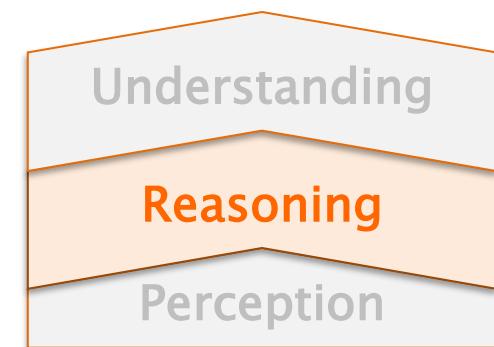
VISUAL REASONING



Graph layout

Layout and visual attributes allow:

- **Discrimination**
 - ◆ Distinguish visual objects or group of –
- **Comparison**
 - ◆ Place visual objects in order
- **Magnitude assessment**
 - ◆ Evaluate the (relative) magnitude of visual objects

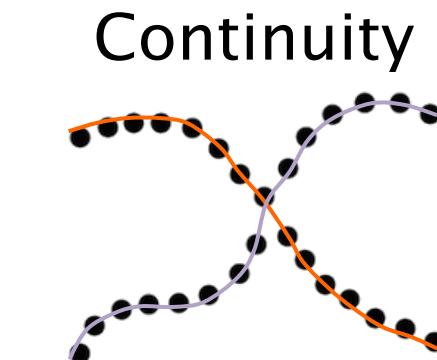
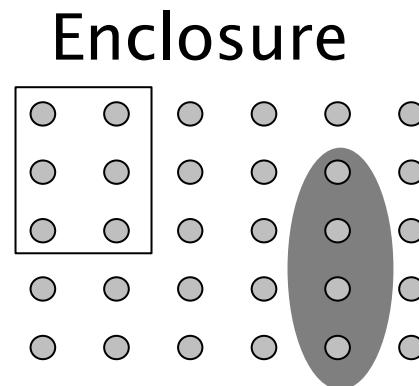
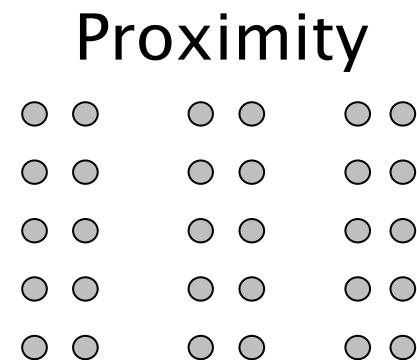
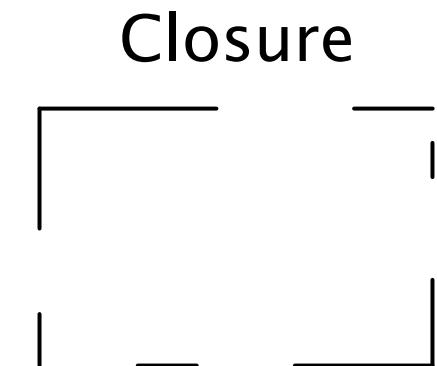
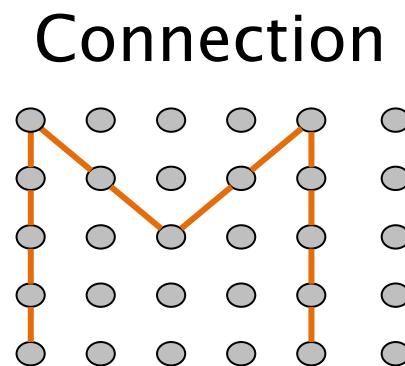
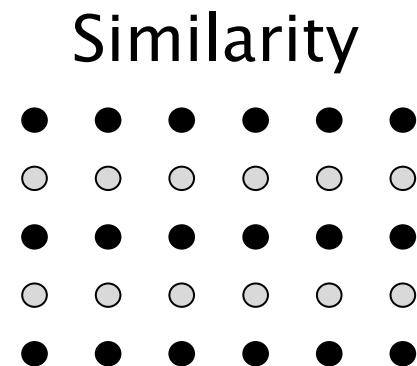


Gestalt principles

- Visual features that lead us to group visual objects together
 - ◆ Proximity
 - ◆ Similarity
 - ◆ Enclosure
 - ◆ Closure
 - ◆ Continuity
 - ◆ Connection

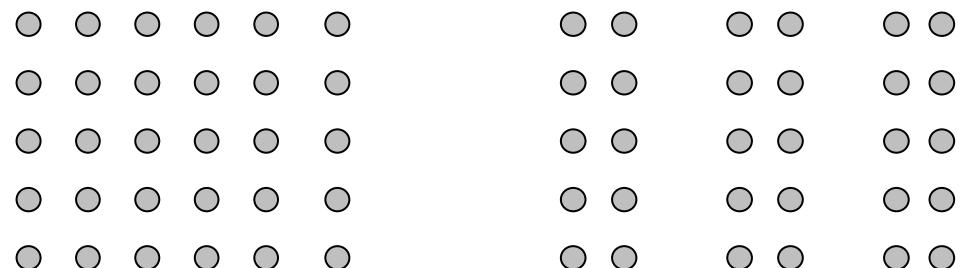
Gestalt principles

- Visual features that lead the viewer to group visual objects together



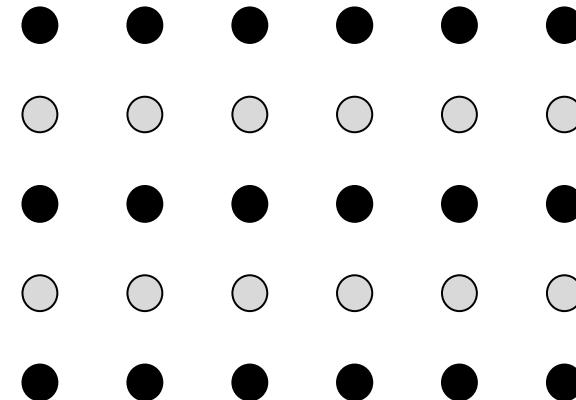
Gestalt principles

- Visual attributes/patterns that lead observer to group objects together
 - ◆ **Proximity**
 - ◆ Similarity
 - ◆ Enclosure
 - ◆ Closure
 - ◆ Continuity
 - ◆ Connection



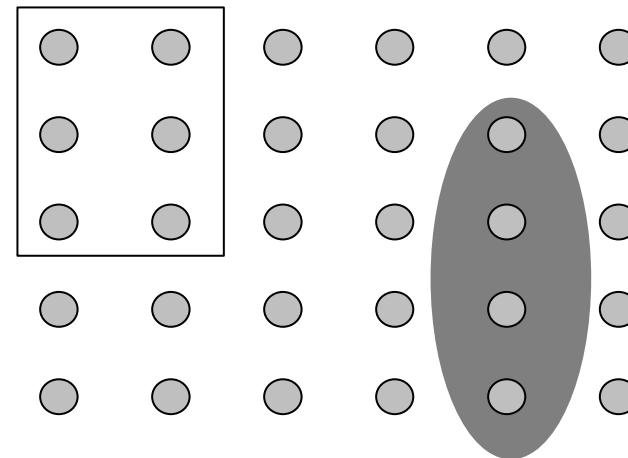
Gestalt principles

- Visual attributes/patterns that lead observer to group objects together
 - ◆ Proximity
 - ◆ **Similarity**
 - ◆ Enclosure
 - ◆ Closure
 - ◆ Continuity
 - ◆ Connection



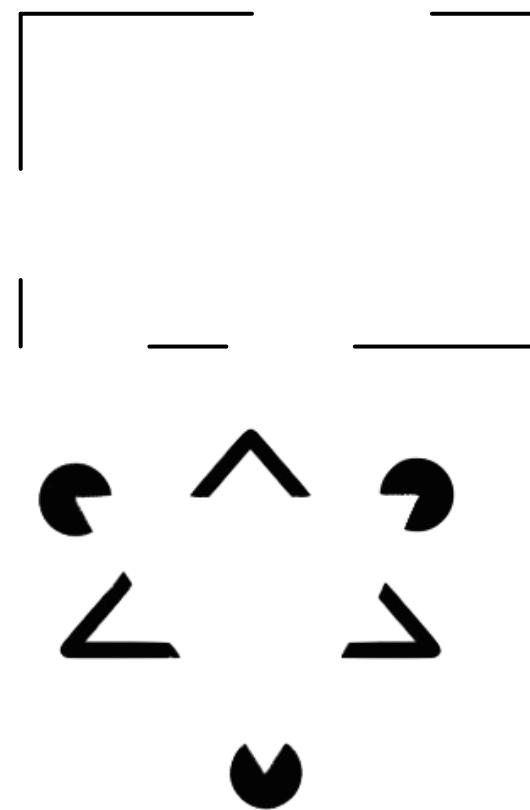
Gestalt principles

- Visual attributes/patterns that lead observer to group objects together
 - ◆ Proximity
 - ◆ Similarity
 - ◆ **Enclosure**
 - ◆ Closure
 - ◆ Continuity
 - ◆ Connection



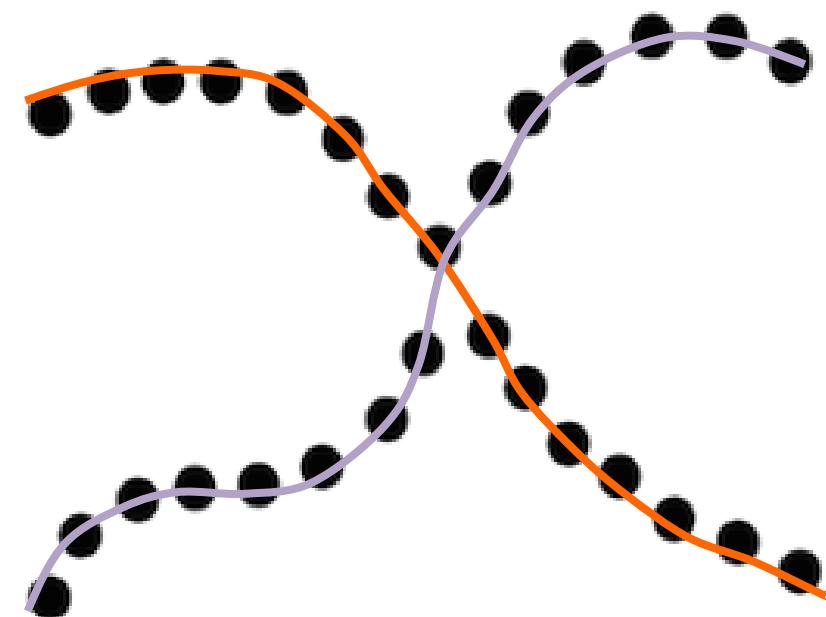
Gestalt principles

- Visual attributes/patterns that lead observer to group objects together
 - ◆ Proximity
 - ◆ Similarity
 - ◆ Enclosure
 - ◆ **Closure**
 - ◆ Continuity
 - ◆ Connection



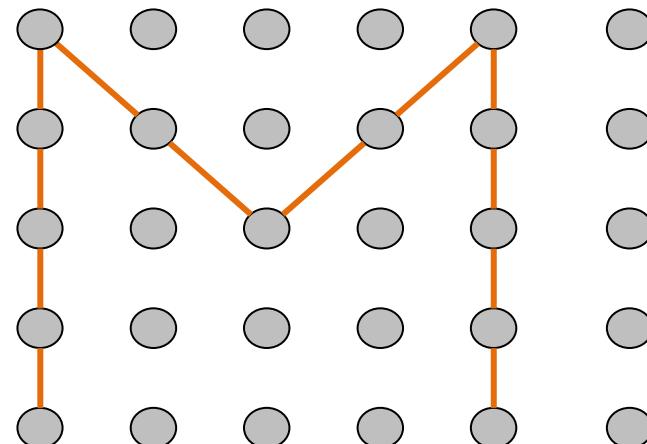
Gestalt principles

- Visual attributes/patterns that lead observer to group objects together
 - ◆ Proximity
 - ◆ Similarity
 - ◆ Enclosure
 - ◆ Closure
 - ◆ **Continuity**
 - ◆ Connection

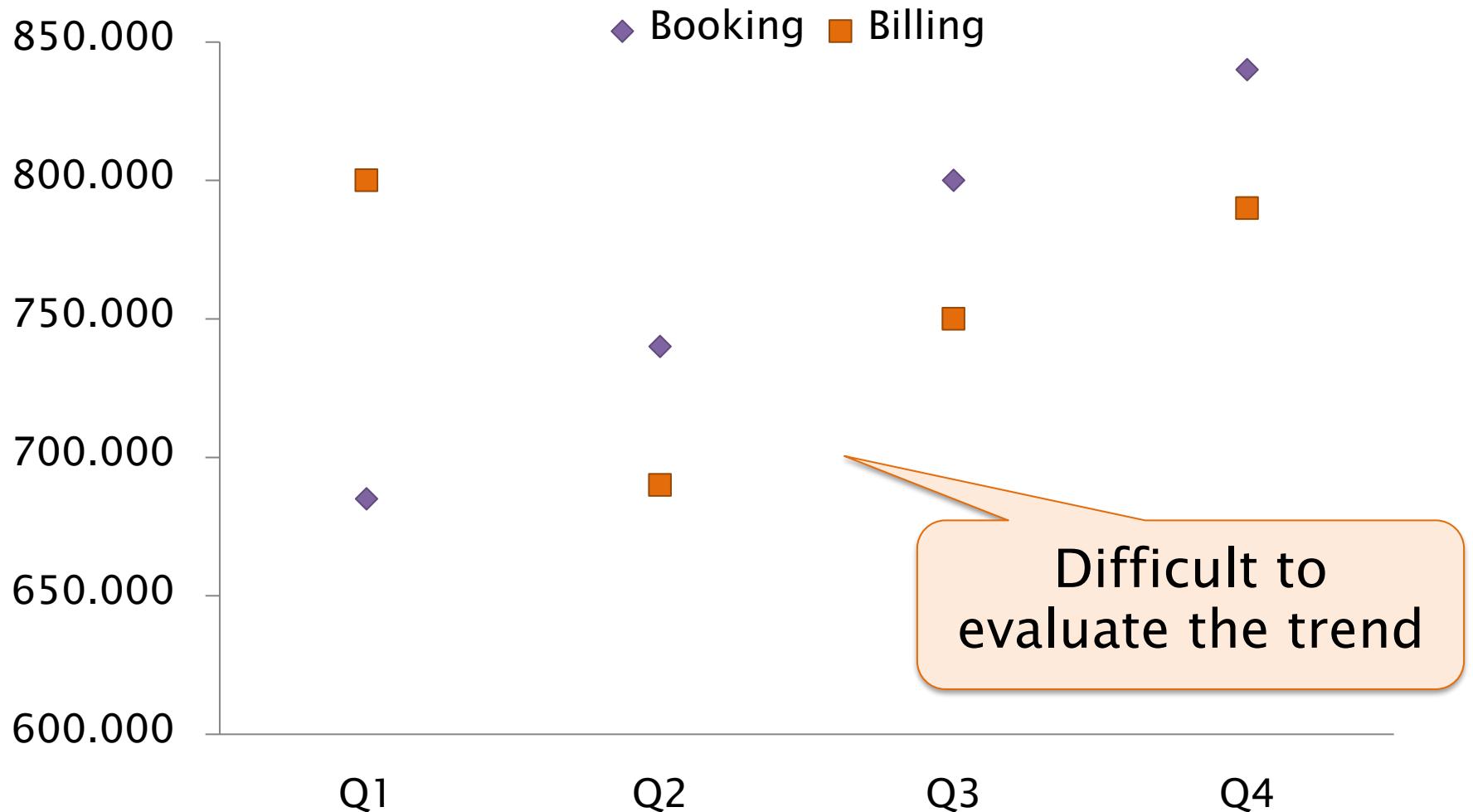


Gestalt principles

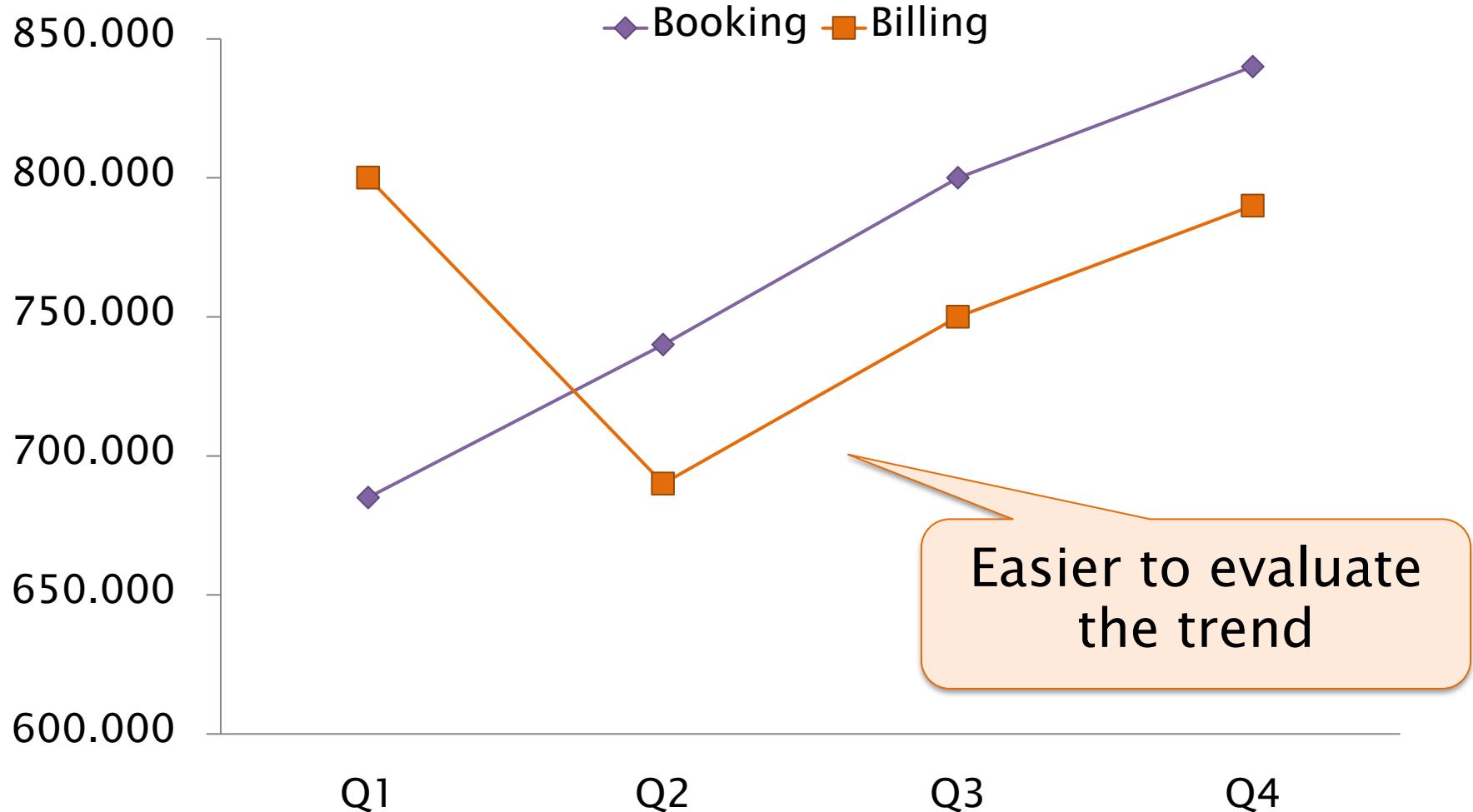
- Visual attributes/patterns that lead observer to group objects together
 - ◆ Proximity
 - ◆ Similarity
 - ◆ Enclosure
 - ◆ Closure
 - ◆ Continuity
 - ◆ Connection



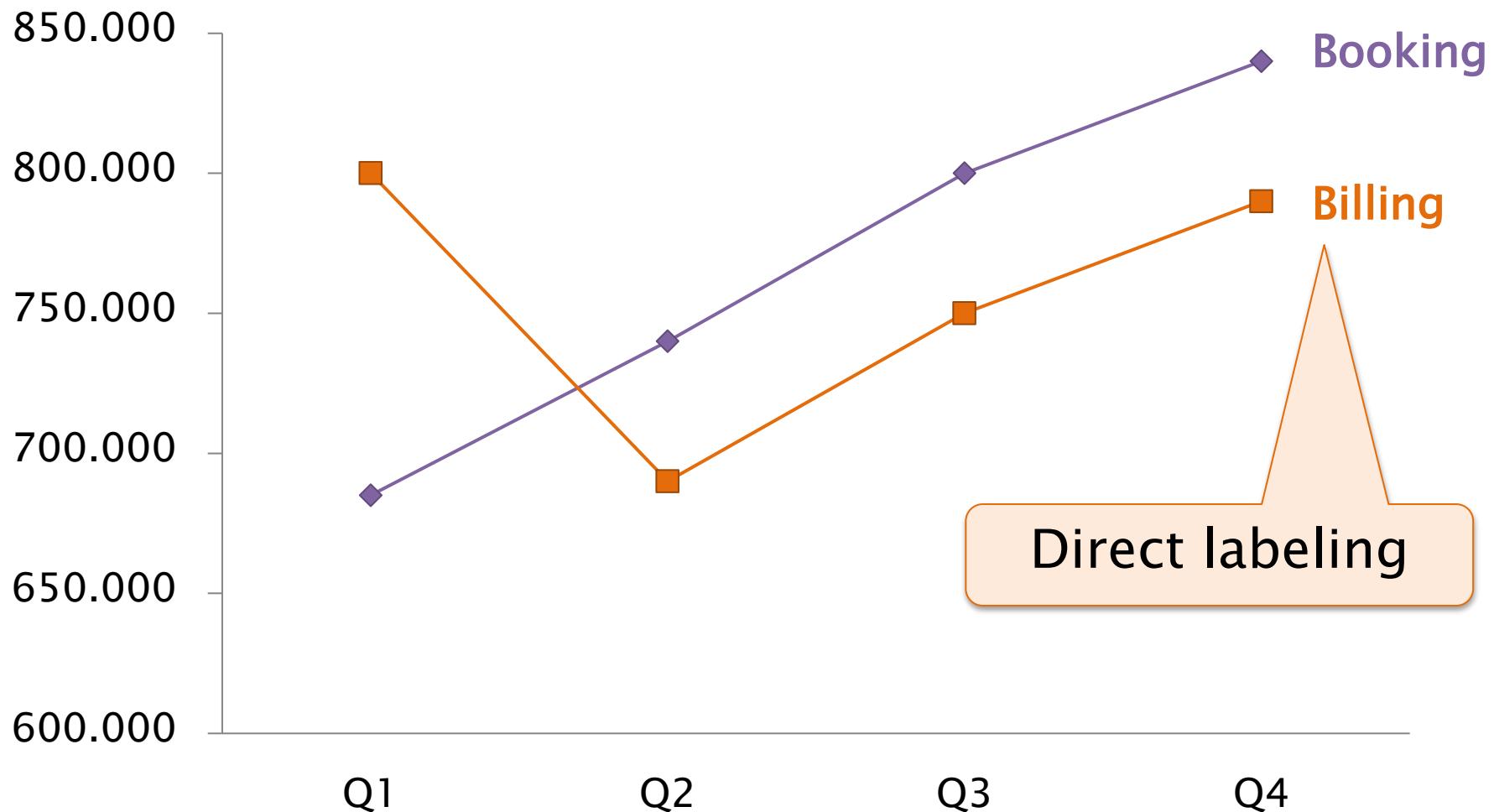
Similarity in Shape & Color



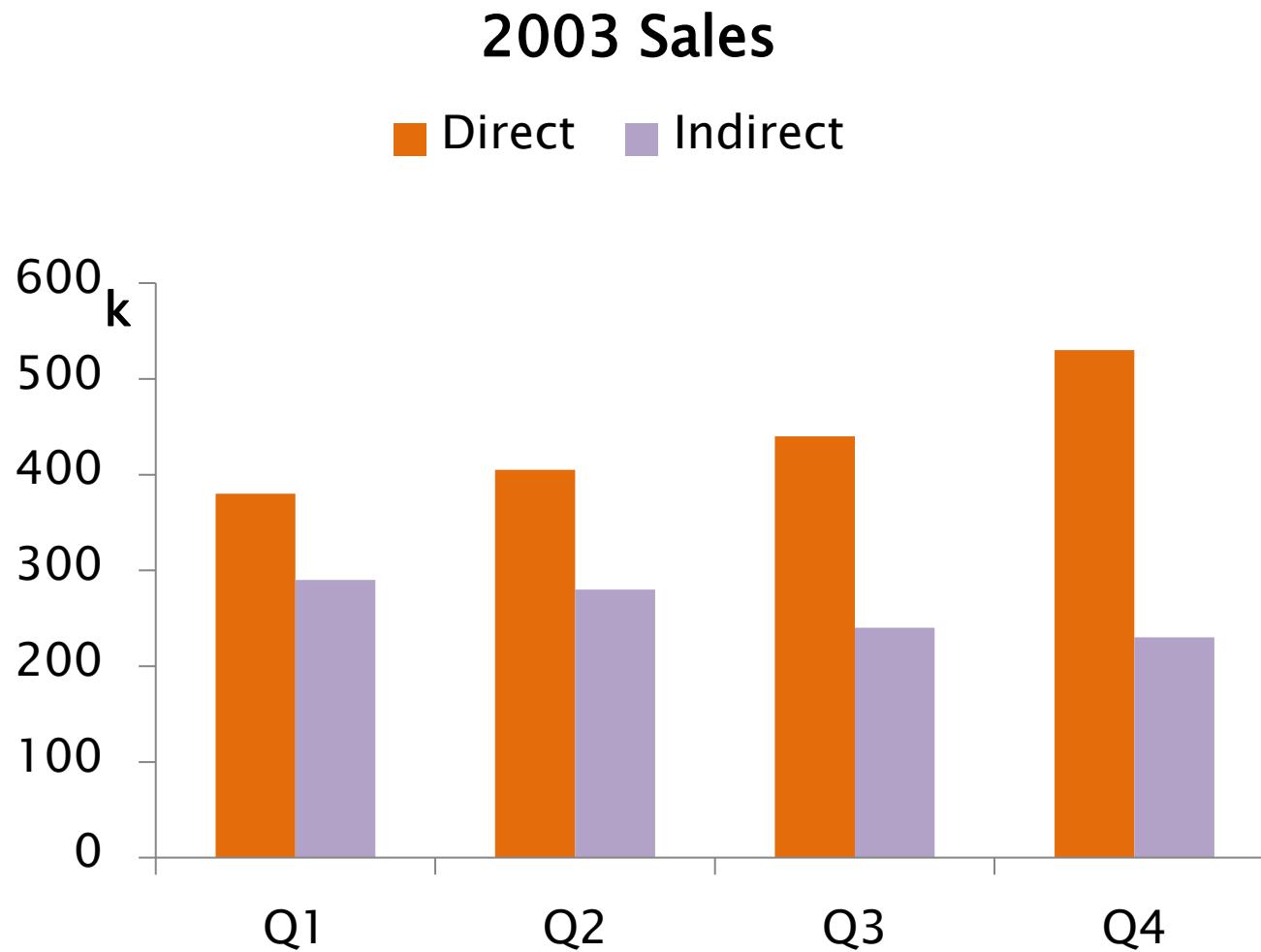
Similarity+Connection



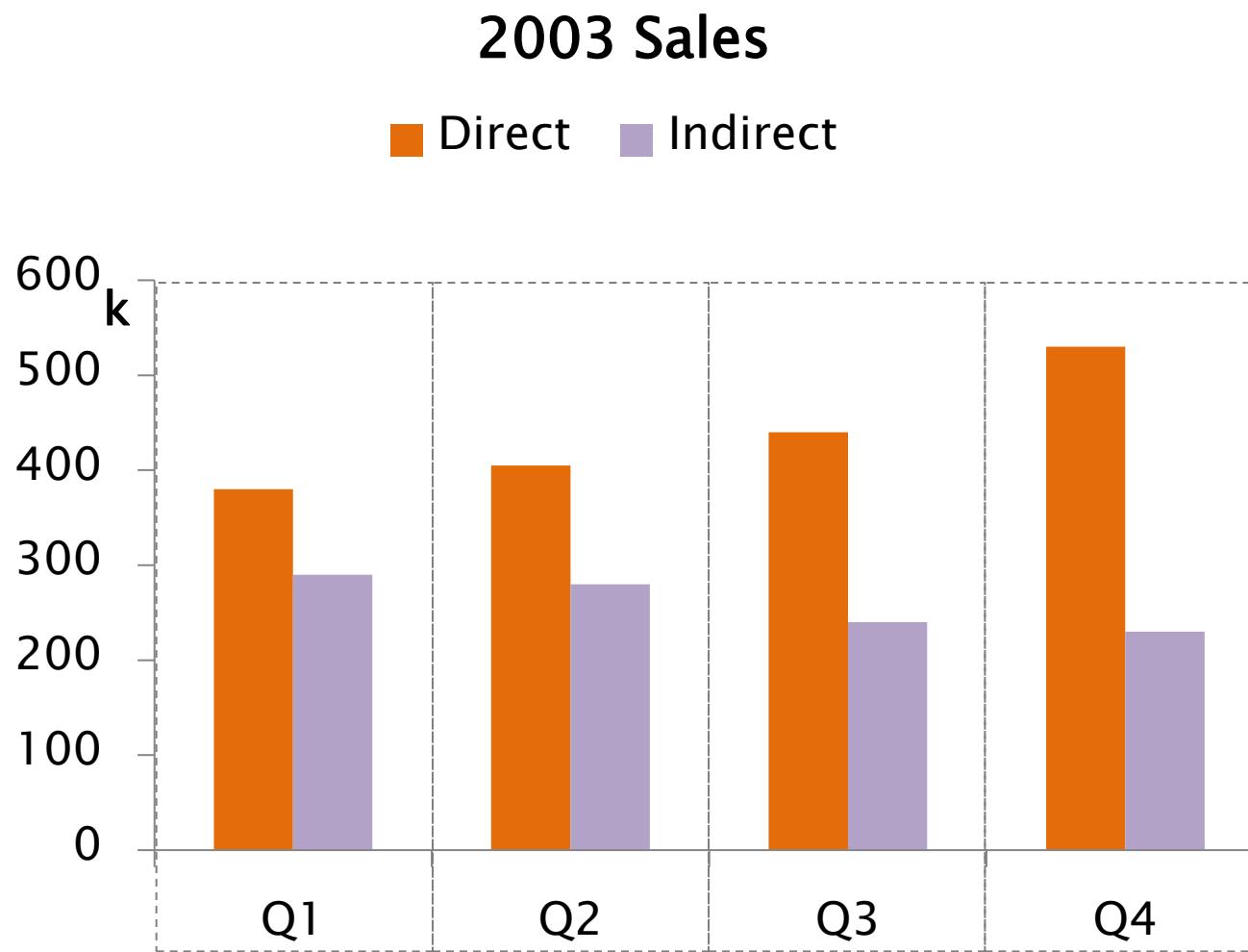
Similarity+Connection+Proximity



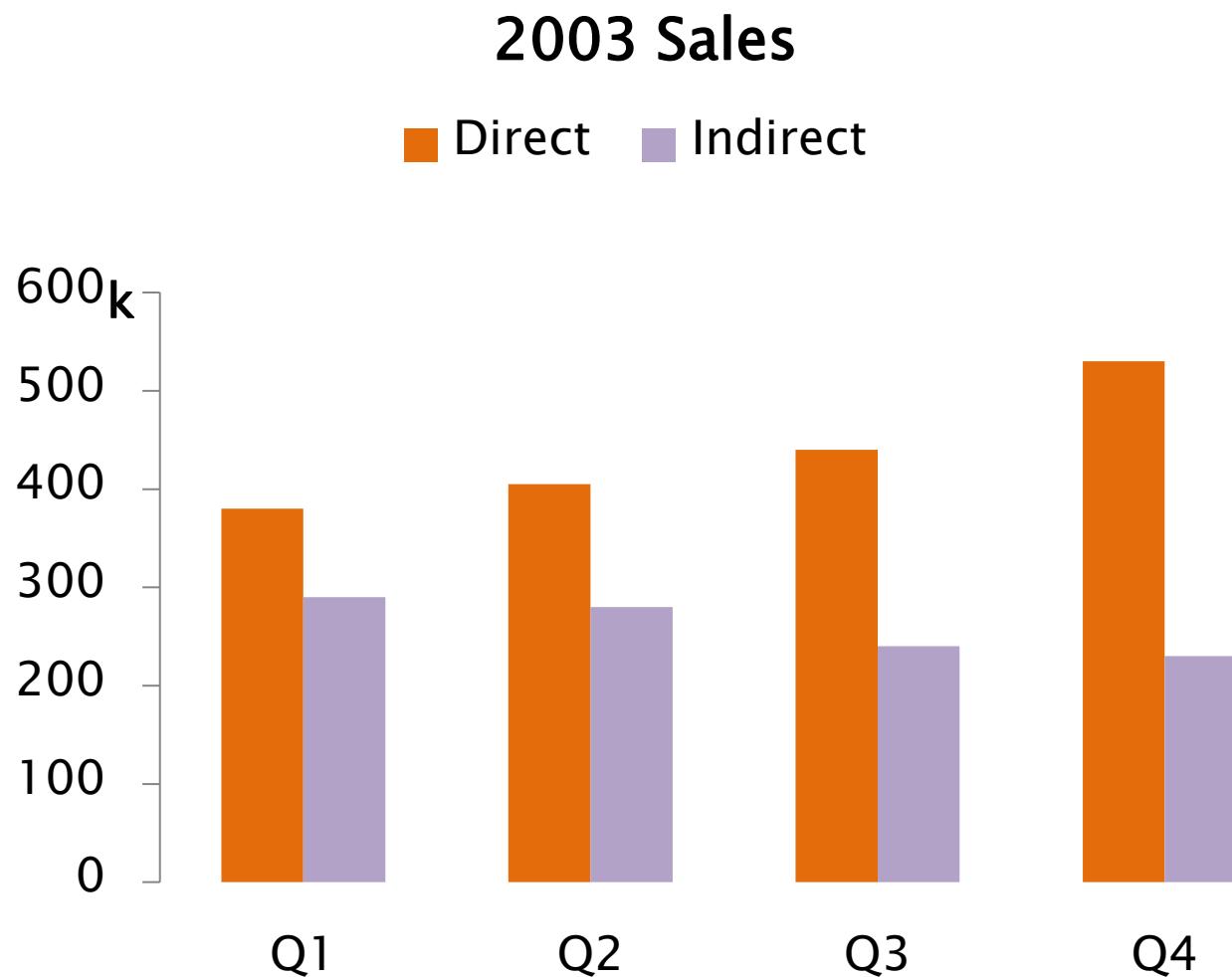
Similarity × Proximity



Similarity × Proximity & Enclosure



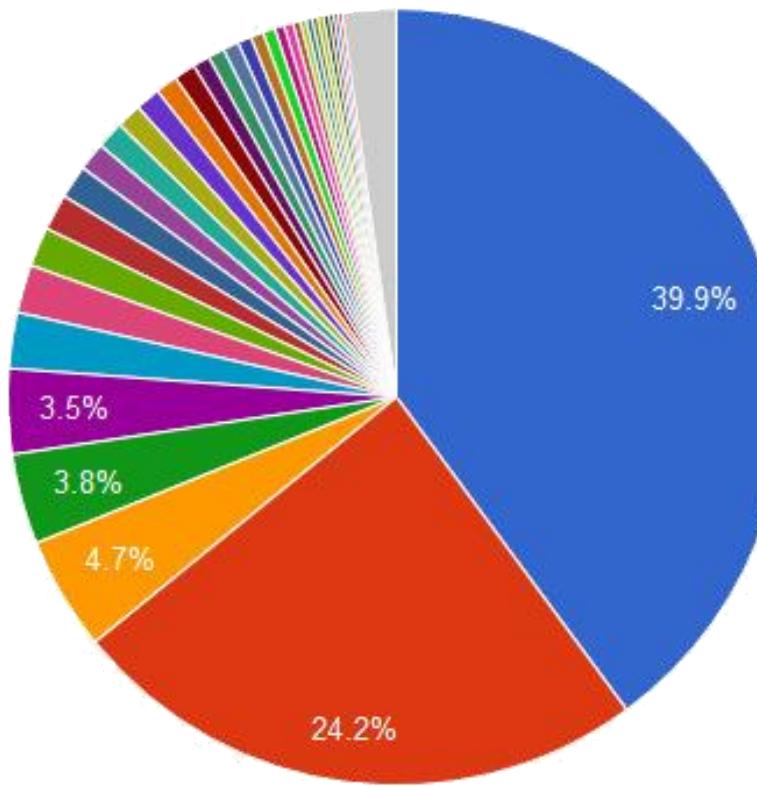
Continuity replaces axis



Distinct perceptions

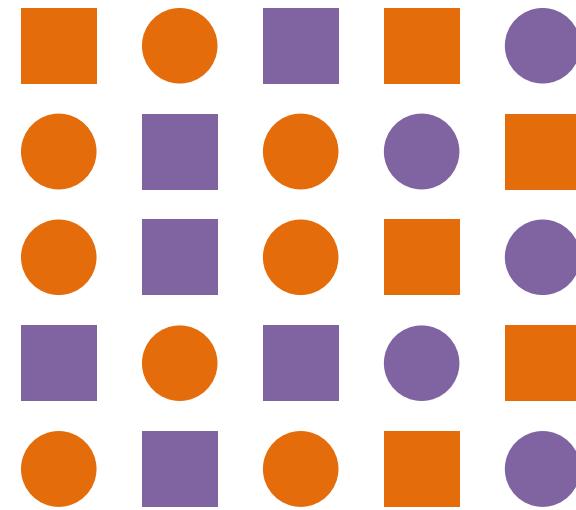
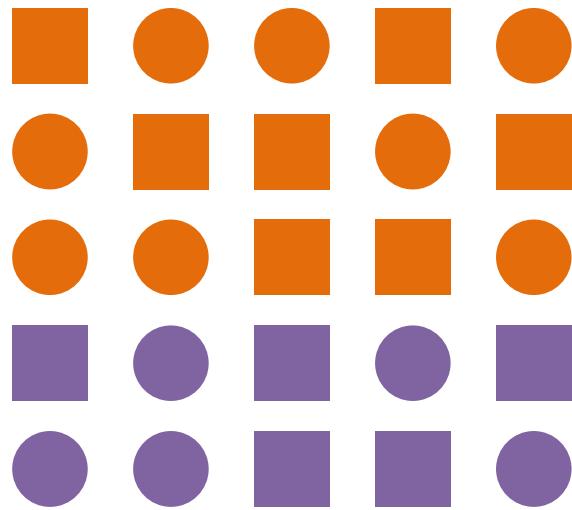
- The immediacy of any pre-attentive cue declines as the variety of alternative patterns increases
 - ◆ Even if all the distracting patterns are individually distinct from the target
 - ◆ For each single attribute no more than **four** distinct levels are discernible

Rainbow Pies

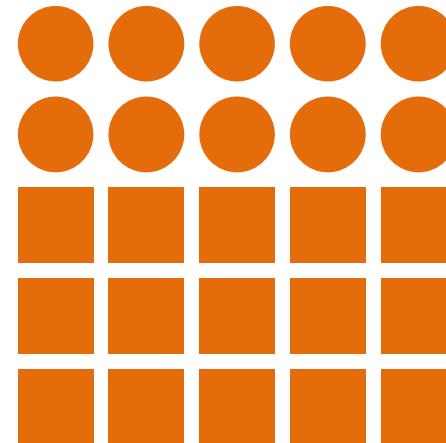
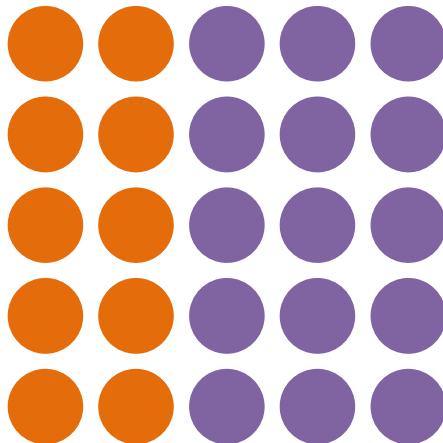
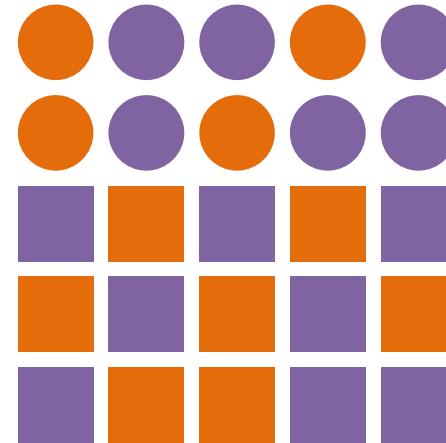
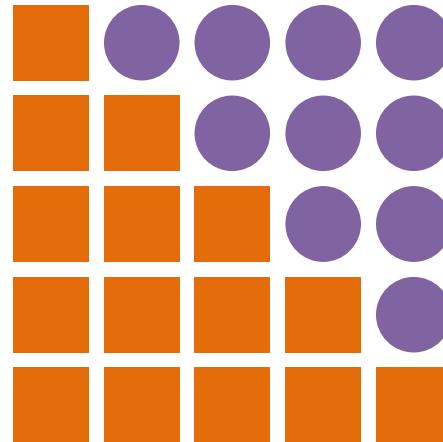
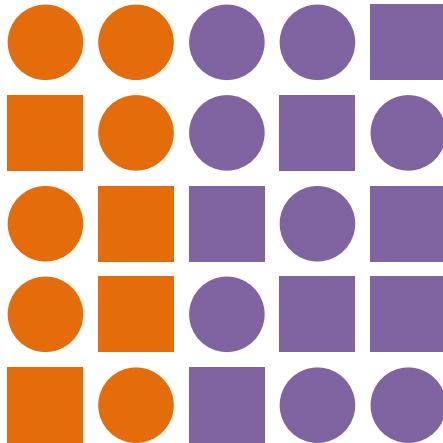


- Google Analytics
- Google Universal Analytics
- Quantcast
- comScore
- Yandex Metrica
- LiveInternet
- Moat
- Nielsen NetRatings
- Adometry (Google)
- CrazyEgg
- Histats
- Drawbridge
- Piwik
- StatCounter
- Yahoo Analytics
- Convertro (AOL)
- whos.amung.us
- Omniture (Adobe)
- Hotjar
- Chartbeat

Attribute Interference



Attribute Interference



Cultural conventions

- Reading proceed from left to right and from top to bottom
 - ◆ At least in western culture
- What is at the top (on the left) precedes what is at the bottom (on the right) in terms of
 - ◆ Importance
 - ◆ Ordering
 - ◆ Time

Emphasis

Attribute

Line width

Size

Color intensity

2-D position

Tables

Boldface text

Bigger tables
Larger fonts

Darker or brighter colors

Positioned at the top
Positioned at the left
Positioned in the center

Graphs

Thicker lines

Bigger graphs
Wider bars
Bigger symbols

References

- C. Ware. *Information Visualization: Perception for Design*. Morgan Kaufmann Publishers, Inc., San Francisco, California, 2000
- C. Healey, and J. Enns. *Attention and Visual Memory in Visualization and Computer Graphics*. IEEE Transactions on Visualization and Computer Graphics, 18(7), 2012
- I. Inbar, N. Tractinsky and J. Meyer.
Minimalism in information visualization:
attitudes towards maximizing the data–ink ratio.
 - ◆ <http://portal.acm.org/citation.cfm?id=1362587>

References

- S.Few, "Practical Rules for Using Color in Charts"
 - ◆ http://www.perceptualedge.com/articles/visual_business_intelligence/rules_for_using_color.pdf
- D. Borland and R. M. Taylor II, "Rainbow Color Map (Still) Considered Harmful," in *IEEE Computer Graphics and Applications*, vol. 27, no. 2, pp. 14-17, March-April 2007.
 - ◆ http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4118486
- <http://www.color-blindness.com>
- <http://www.csc.ncsu.edu/faculty/healey/PP/index.html>

Graph Construction

Data Management and Visualization



SoftEng
<http://softeng.polito.it>

Version 2.6.2
© Marco Torchiano, 2021





This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

To view a copy of this license, visit

<http://creativecommons.org/licenses/by-nc-nd/4.0/>.

You are free: to copy, distribute, display, and perform the work

Under the following conditions:



Attribution. You must attribute the work in the manner specified by the author or licensor.



Non-commercial. You may not use this work for commercial purposes.



No Derivative Works. You may not alter, transform, or build upon this work.

- For any reuse or distribution, you must make clear to others the license terms of this work.
- Any of these conditions can be waived if you get permission from the copyright holder.

Your fair use and other rights are in no way affected by the above.

Data Visualization

Understanding

Information Visualization

Visual Patterns, Trends, Exceptions

Quantitative Reasoning

Quantitative Relationship & Comparison

Visual Perception

Visual Properties & Objects

Representation/Encoding



Grammar of Graphics

- Theory behind graphics construction
 - ◆ Separation of data from aesthetic
 - ◆ Definition of common plot/chart elements
 - ◆ Composition of such common elements
- Building a graphic involves
 1. Specification
 2. Assembly
 3. Display

Leland Wilkinson, *The grammar of graphics*

Specification

- **DATA**: a set of data operations that create variables from datasets
 - ◆ Link variables (e.g., *by index* or *id*)
- **TRANS**: variable transformations (e.g., *rank*)
- **SCALE**: scale transformations (e.g., *log*)
- **COORD**: a coordinate system (e.g., *polar*)
- **ELEMENT**: visual objects (e.g., *points*) and their aesthetic attributes (e.g., *color*, *position*)
- **GUIDE**: guides (e.g., *axes*, *legends*)

Specification for a scatter plot

- DATA: $x = x$
- DATA: $y = y$
- TRANS: $x = x$
- TRANS: $y = y$
- SCALE: $\text{linear}(\text{dim}(1))$
- SCALE: $\text{linear}(\text{dim}(2))$
- COORD: $\text{rect}(\text{dim}(1, 2))$
- GUIDE: $\text{axis}(\text{dim}(1))$
- GUIDE: $\text{axis}(\text{dim}(2))$
- ELEMENT: $\text{point}(\text{position}(x^*y))$

Graph visual components

- Data components
 - ◆ Visual objects associated to measures
 - ◆ Visual attributes
- Layout
 - ◆ Positioning rules (e.g. cartesian coord)
- Support components
 - ◆ Axes
 - ◆ Labels
 - ◆ Legends

VISUAL ENCODING

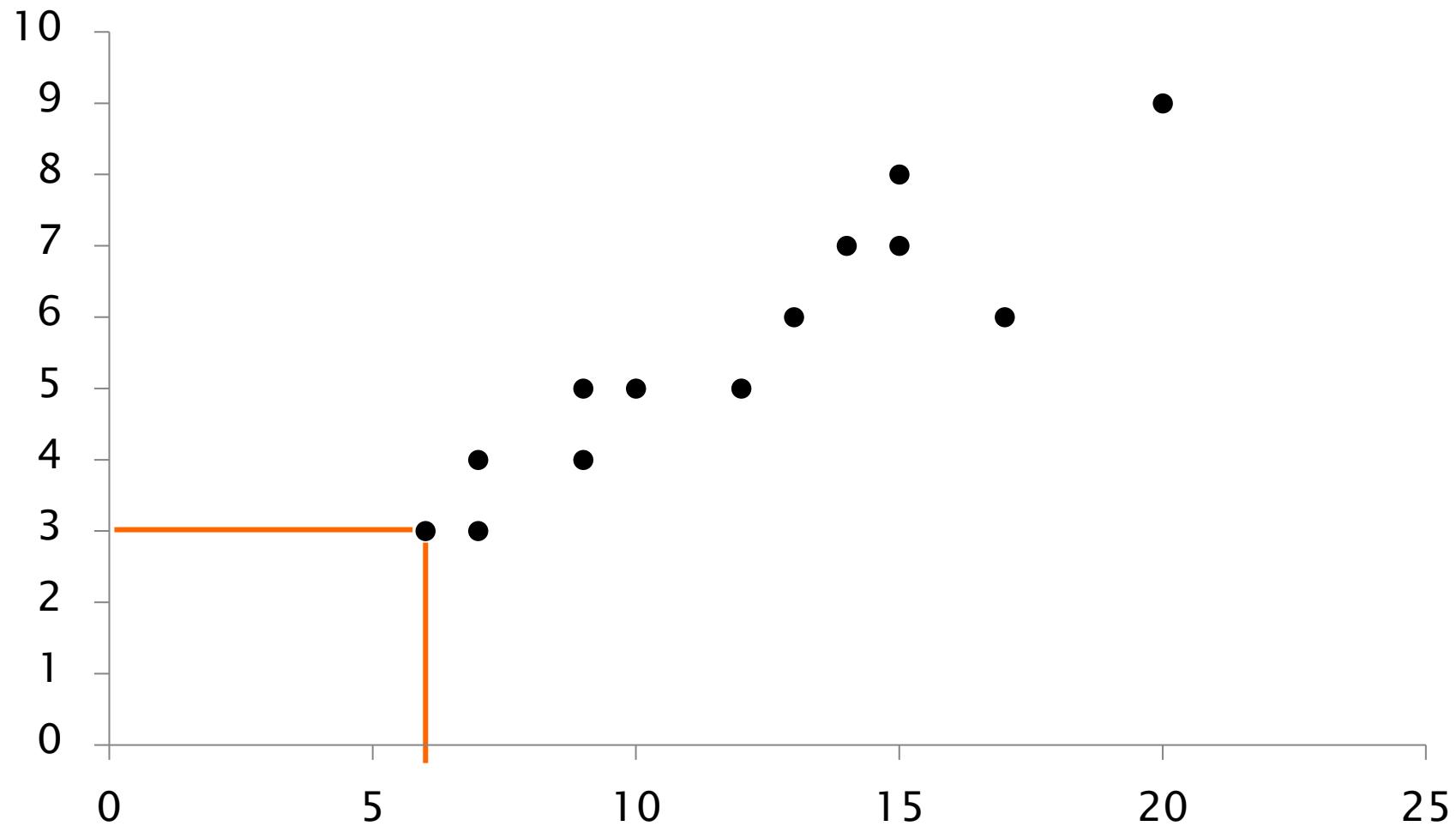
Visual Encoding

- Given a variable (measure), identify:
 - ◆ Visual object
 - ◆ Visual attribute
- Main distinction
 - ◆ Quantitative (interval, ratio, absolute)
 - ◆ Categorical (nominal, ordinal)

Visual Encoding

Object	Attribute
Point	Position (w.r.t. axis/axes)
Line	Length Position (w.r.t. axis/axes) Slope
Bar	Length
Shape	Size (area) Count

Points



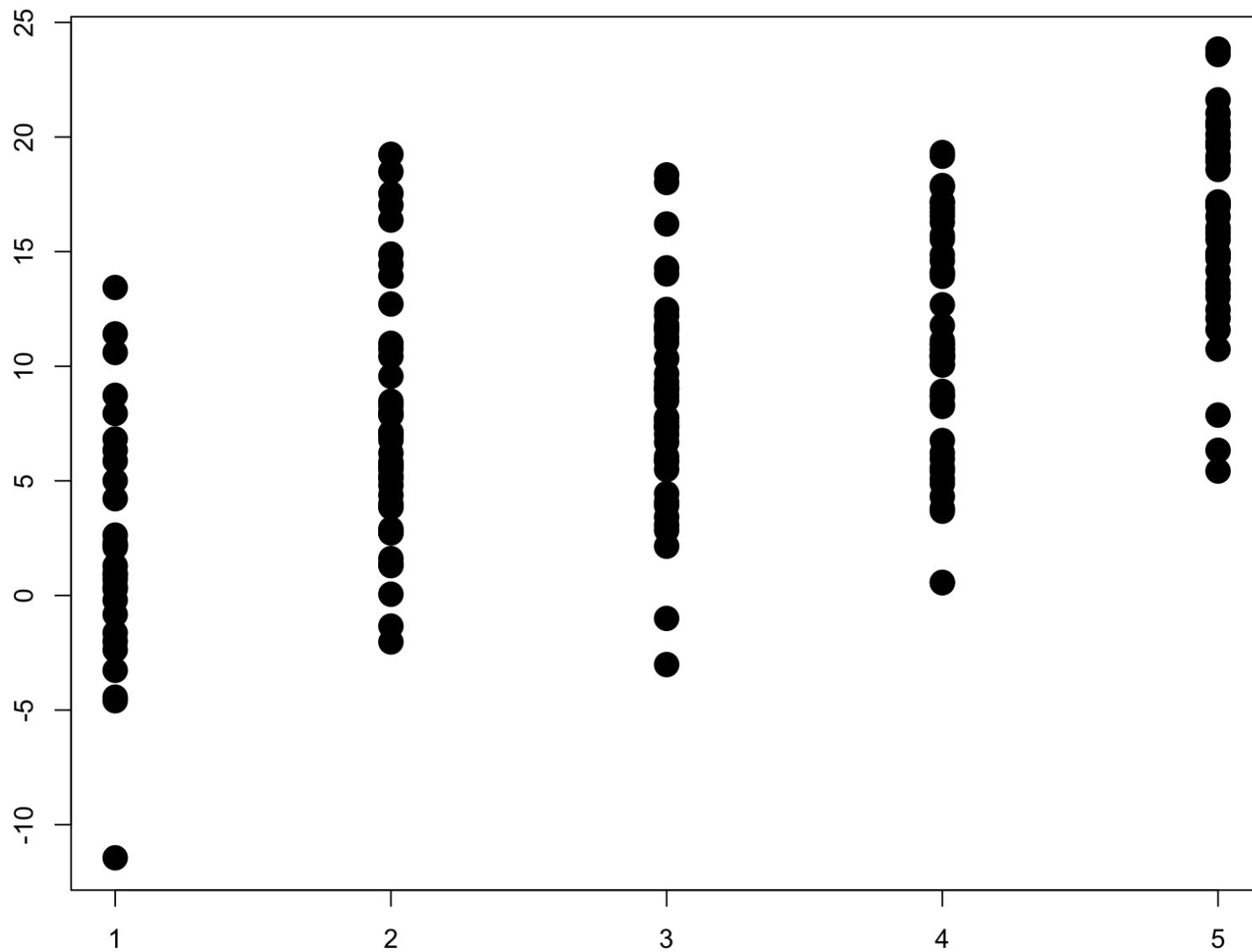
Points Guidelines

- Points must be clearly distinguished
 - ◆ Enlarge points
 - ◆ Select radically distinct shapes (+ ○)
 - ◆ Balance size of points and graph
 - ◆ Use outlined shapes
- Lines must not obscure points

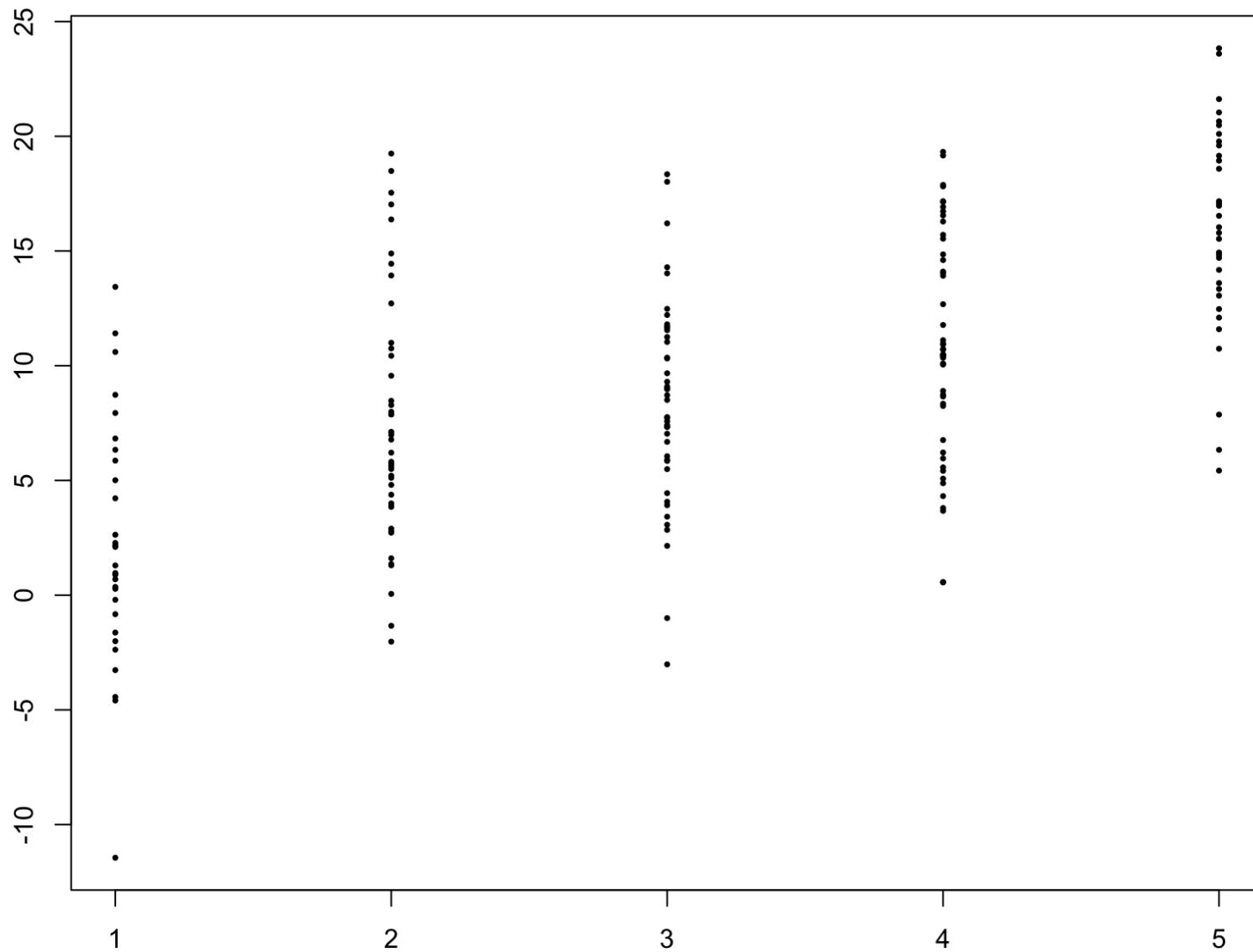
Overplotting

- Phenomenon related to multiple points (or shapes) overlapping
 - ◆ Discrete (integer) measure
 - ◆ Very large dataset
- Solutions
 - ◆ Small shapes
 - ◆ Outlined shapes
 - ◆ Transparent shapes (alpha)
 - ◆ Jittering

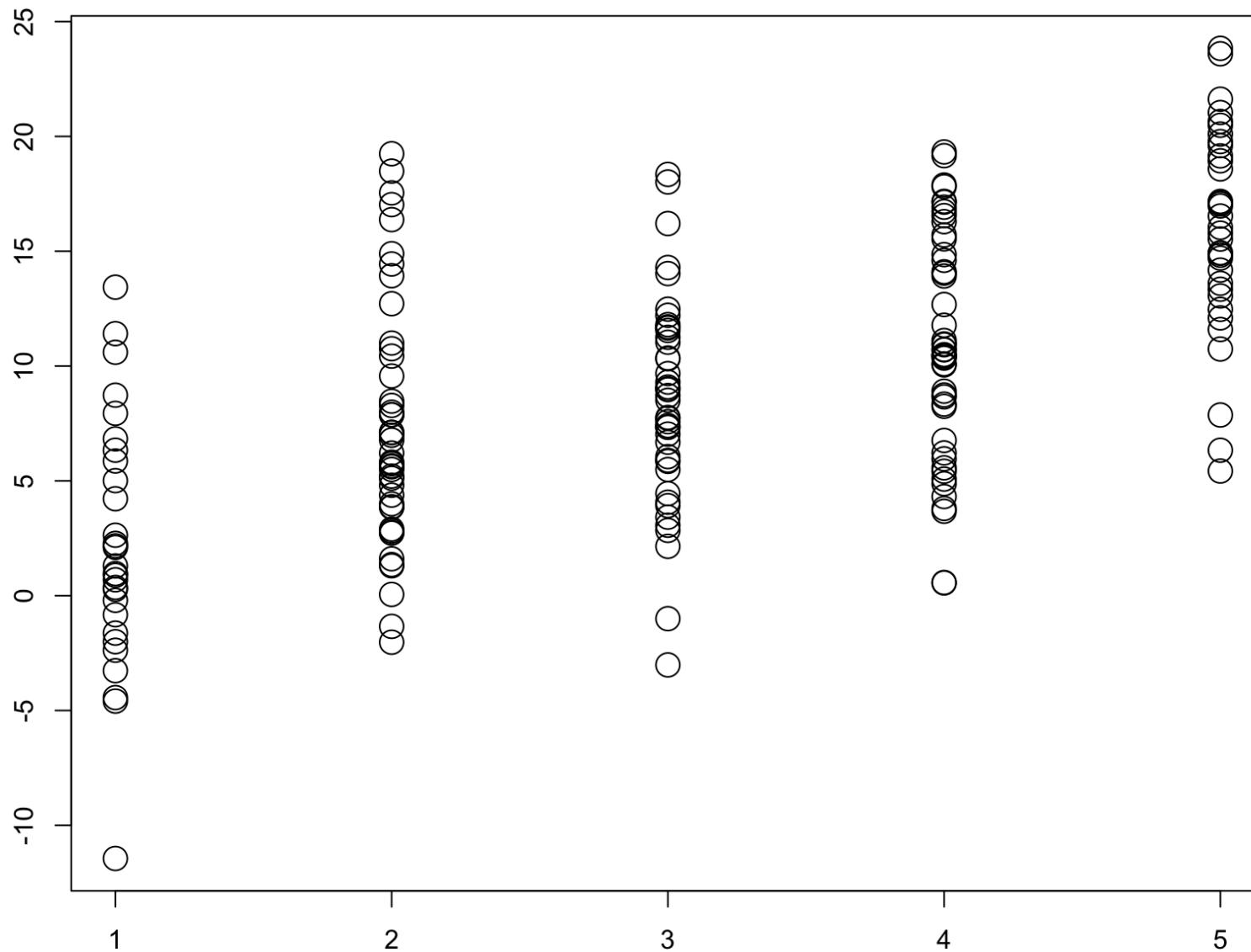
Overplotting example



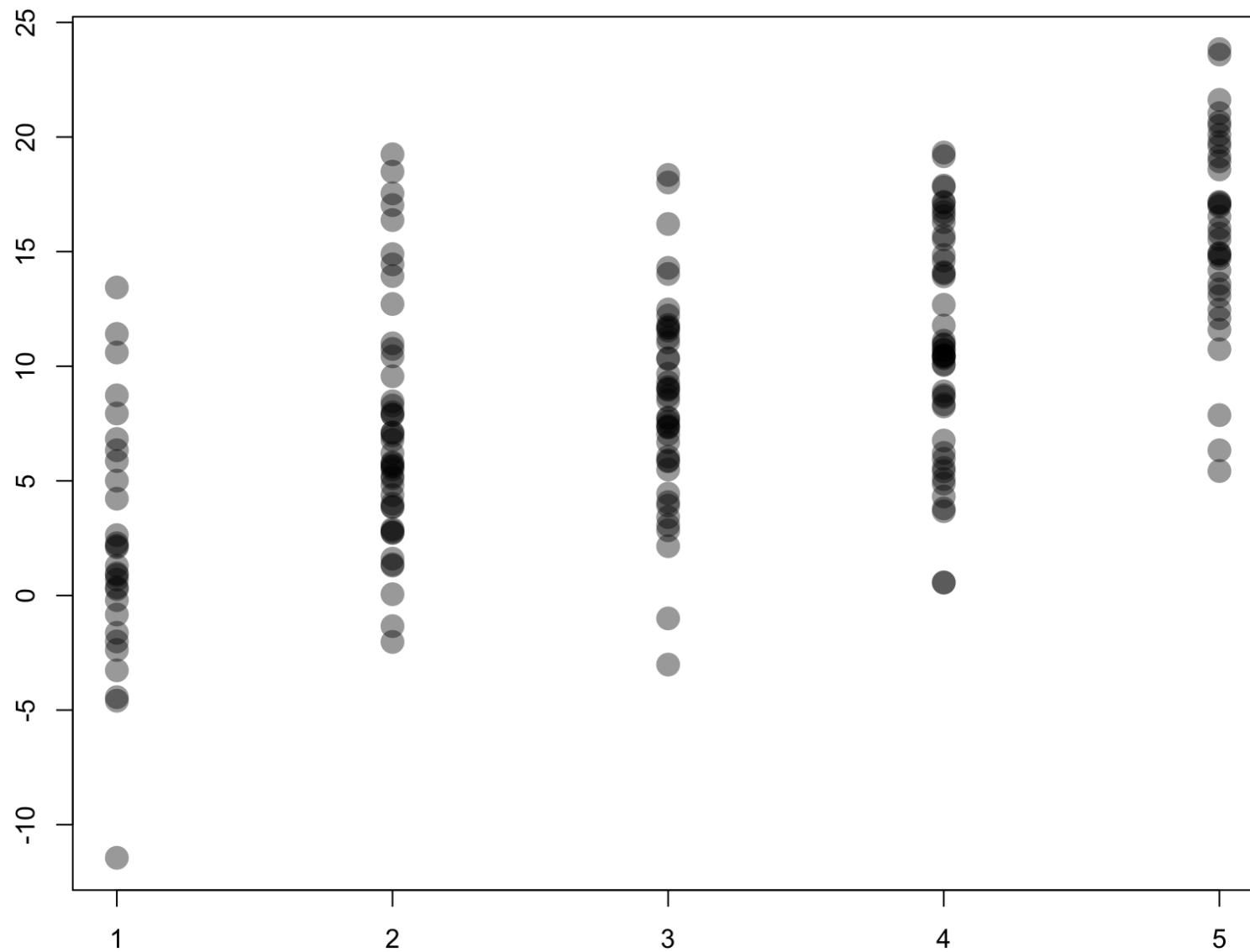
Overplotting – Small



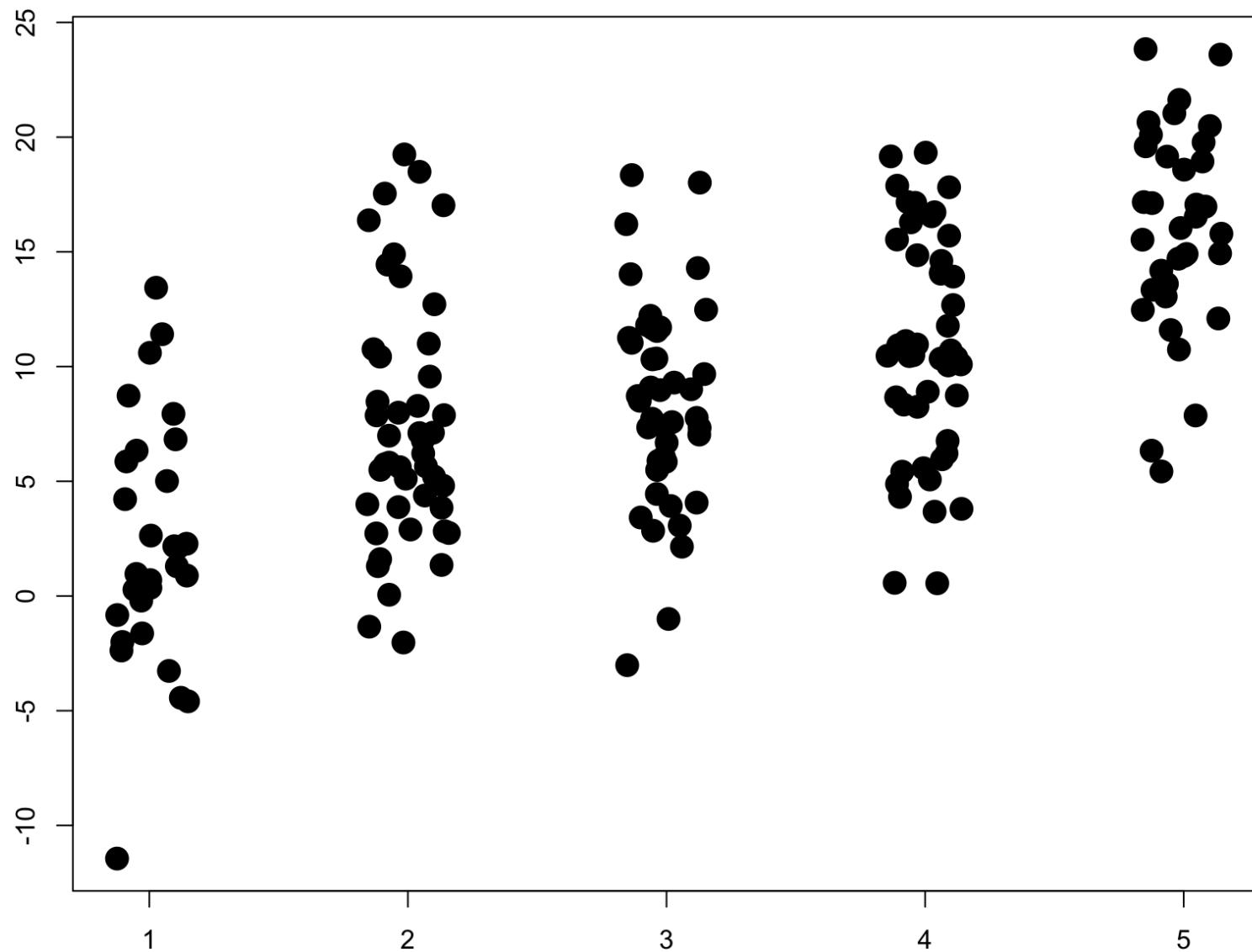
Overplotting - Outlined



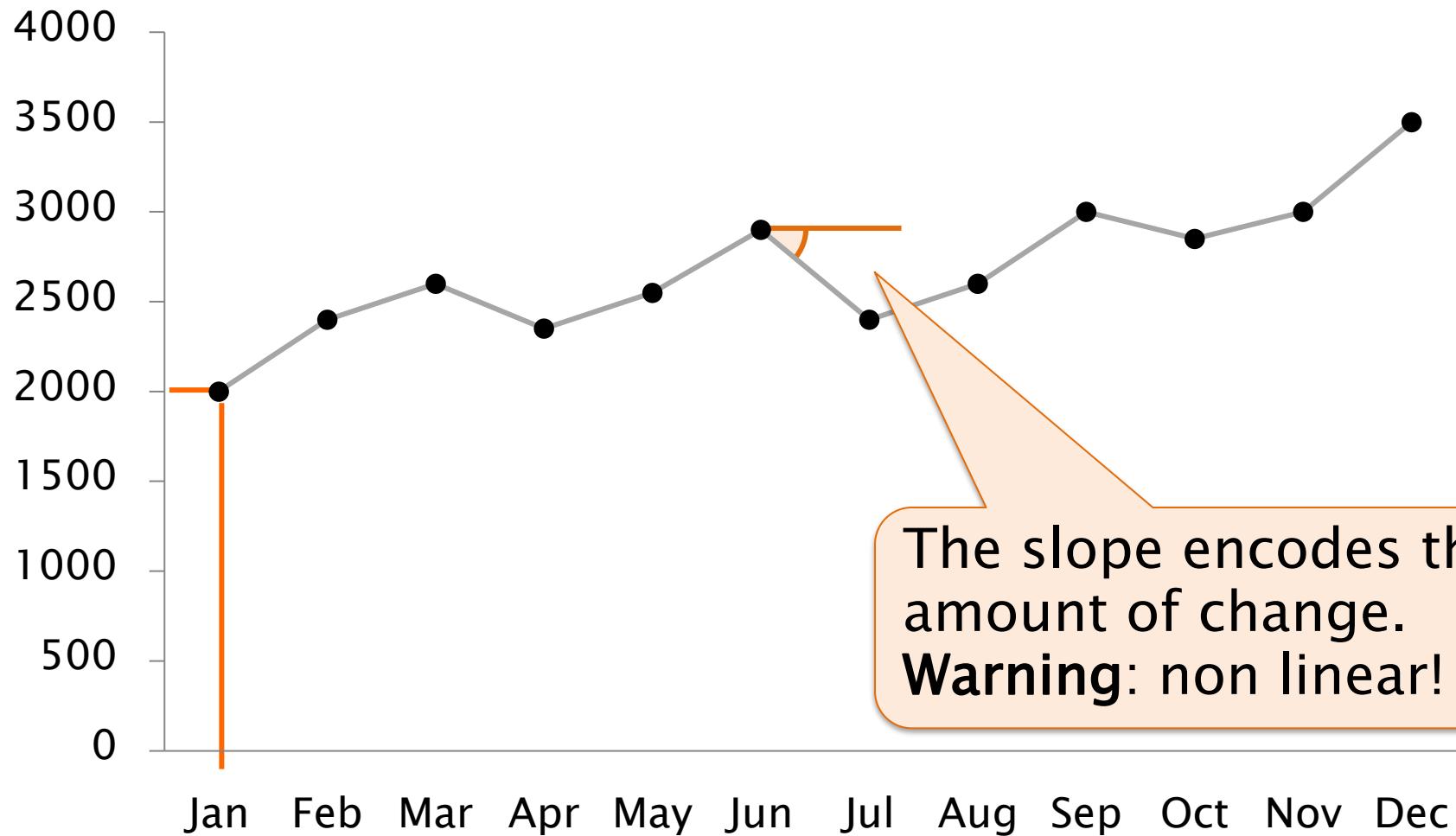
Overplotting – Transparent



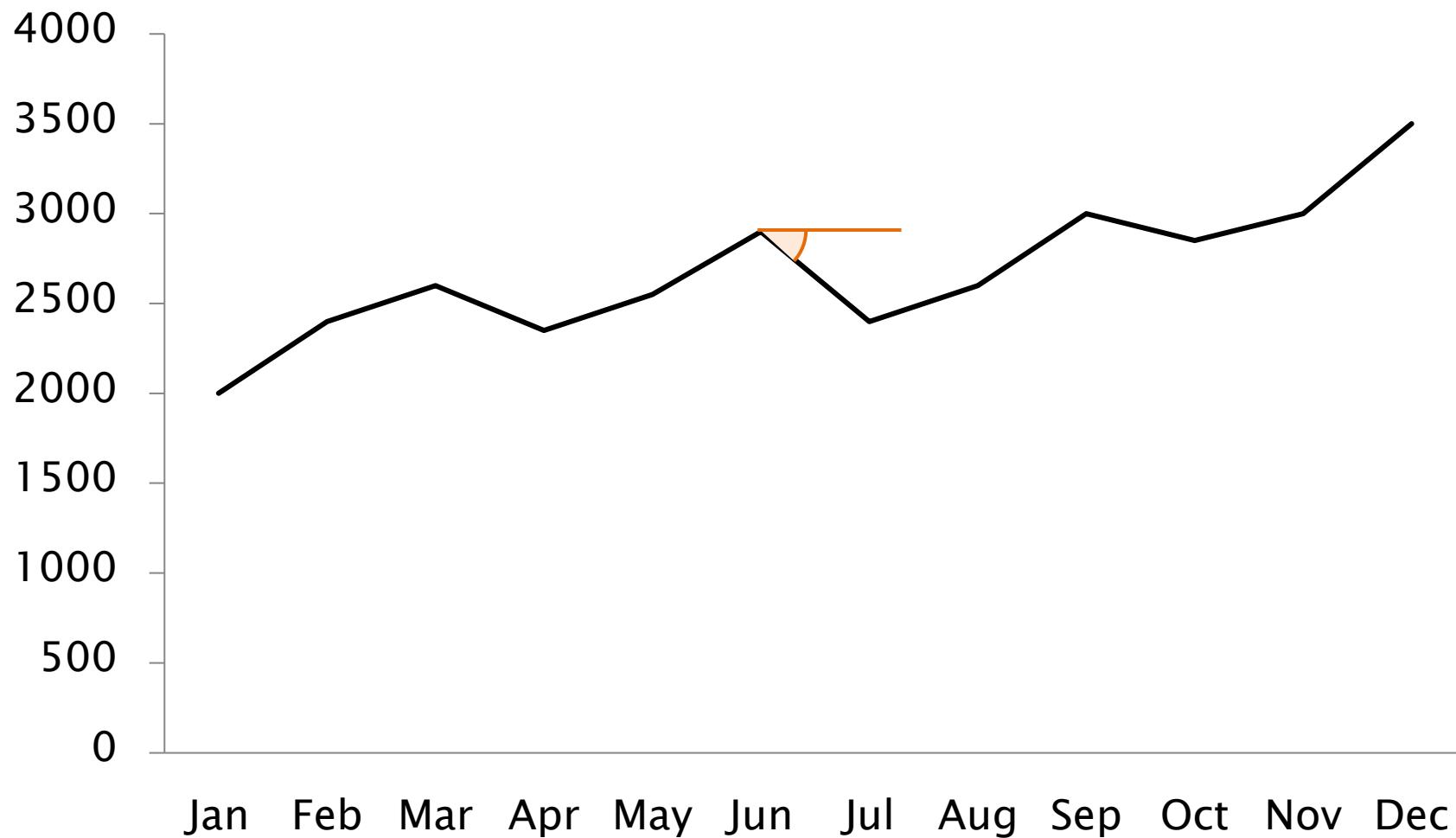
Overplotting – Jittering



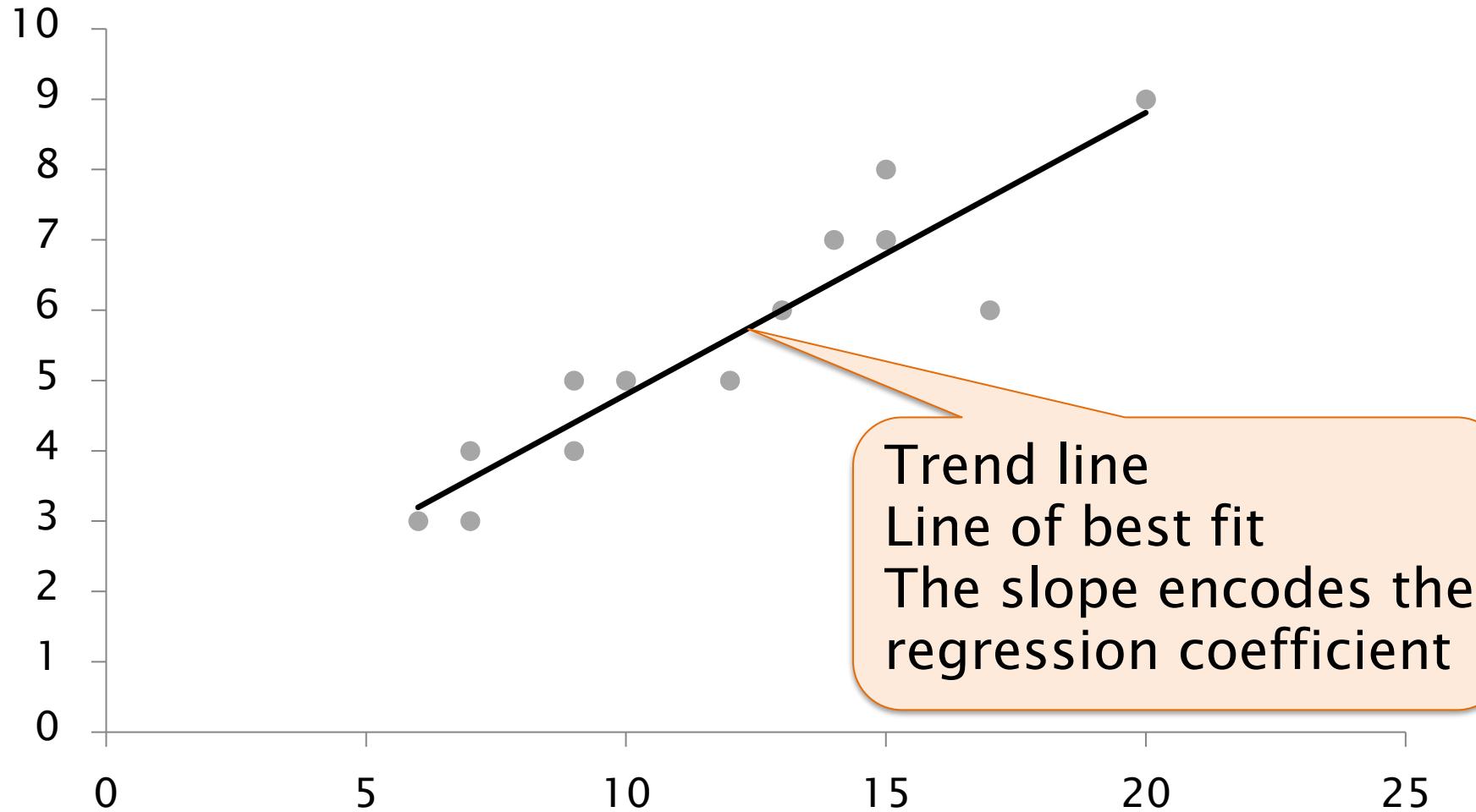
Points and Lines



Slope of lines



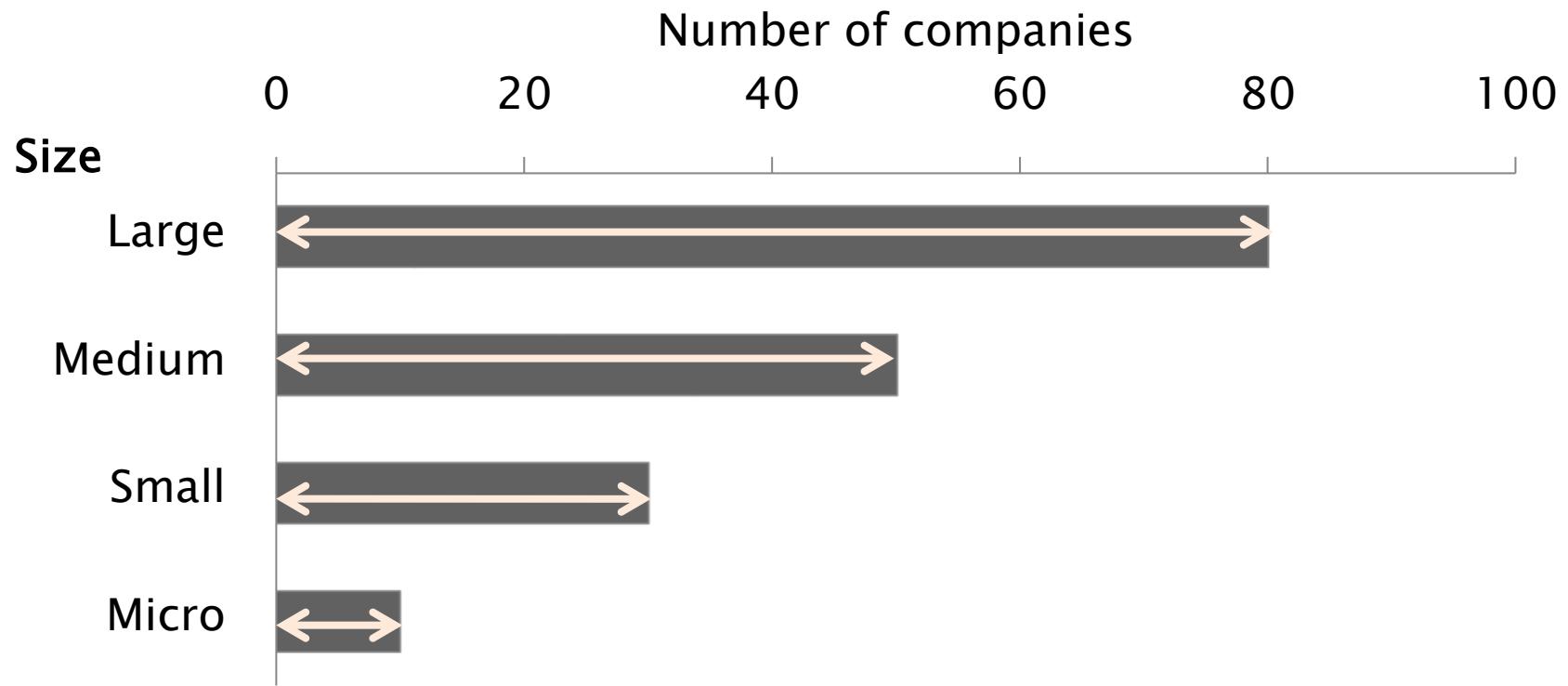
Slope of lines



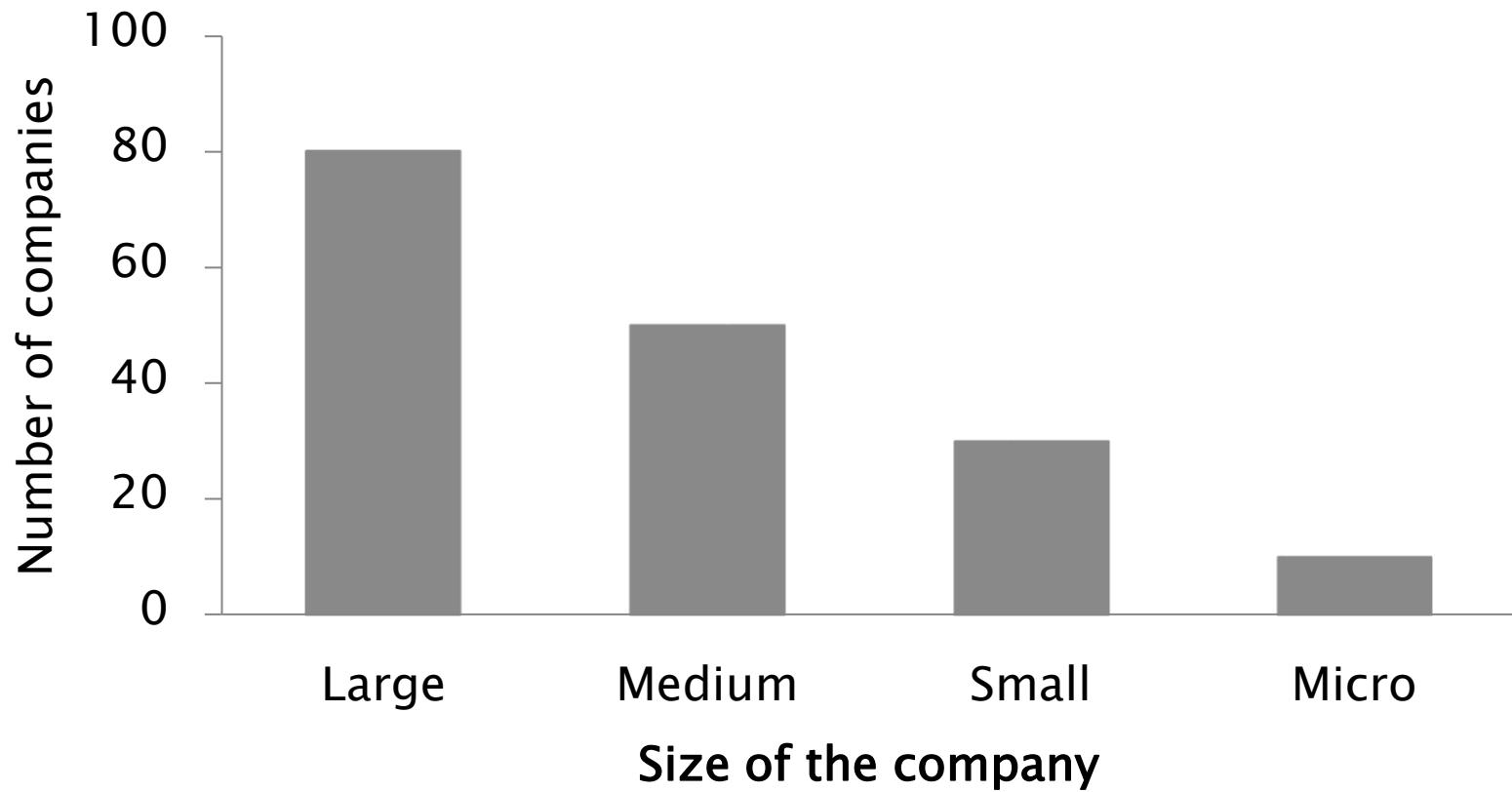
Lines

- Easy perception of trends and overall shape of data
- Best suited for time series
- Variation encoded as slope
 - ◆ Clear direction
 - ◆ Approximate magnitude

Bars (line length)



Vertical Bars (aka Columns)



Bars

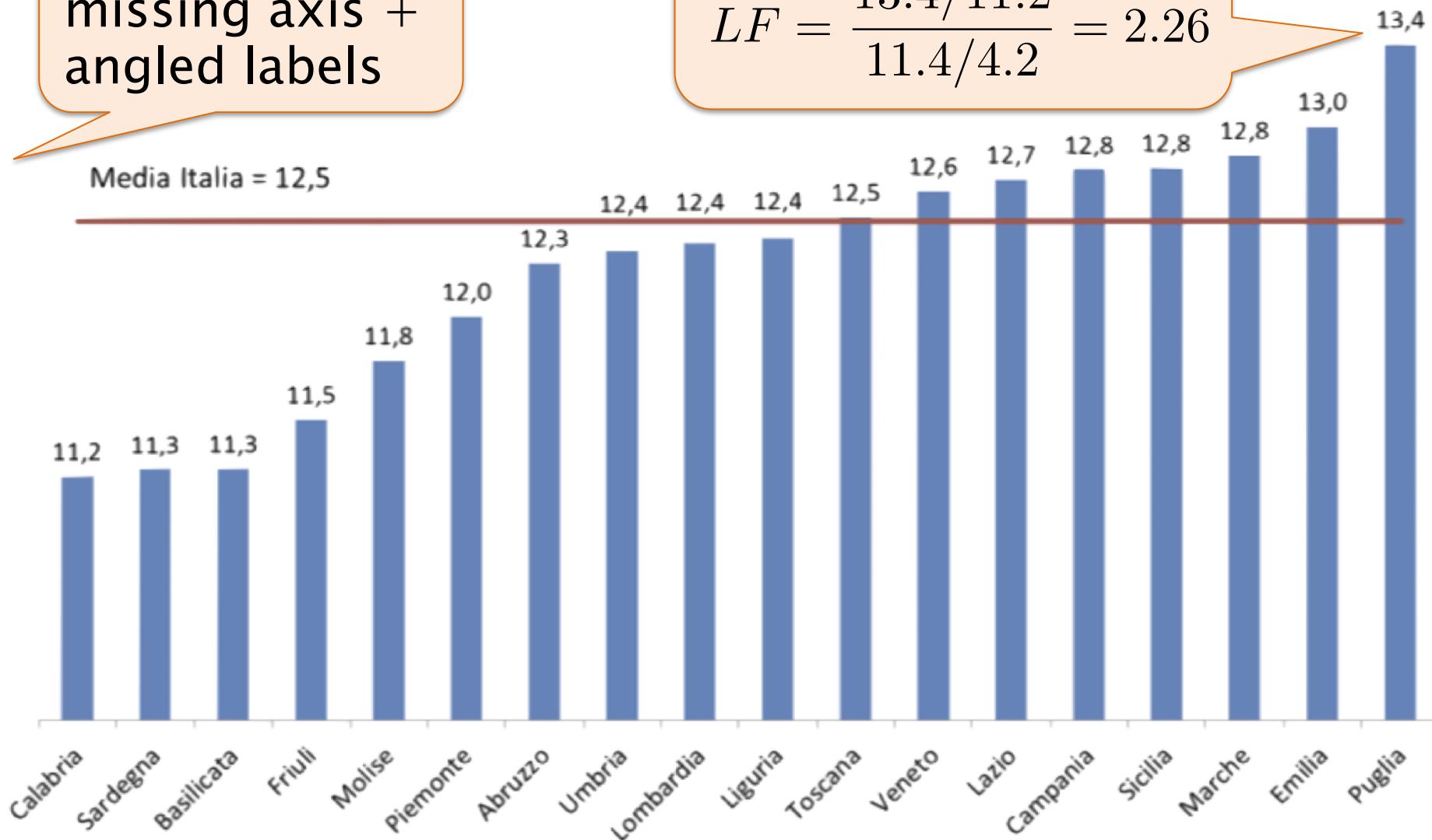
- Quantitative values are encoded only as length of the bars
- Width of bars plays no role
 - ◆ Bars are just very thick lines
- Bars require a zero-based scale
 - ◆ See: Lie factor!

Bar must be zero based

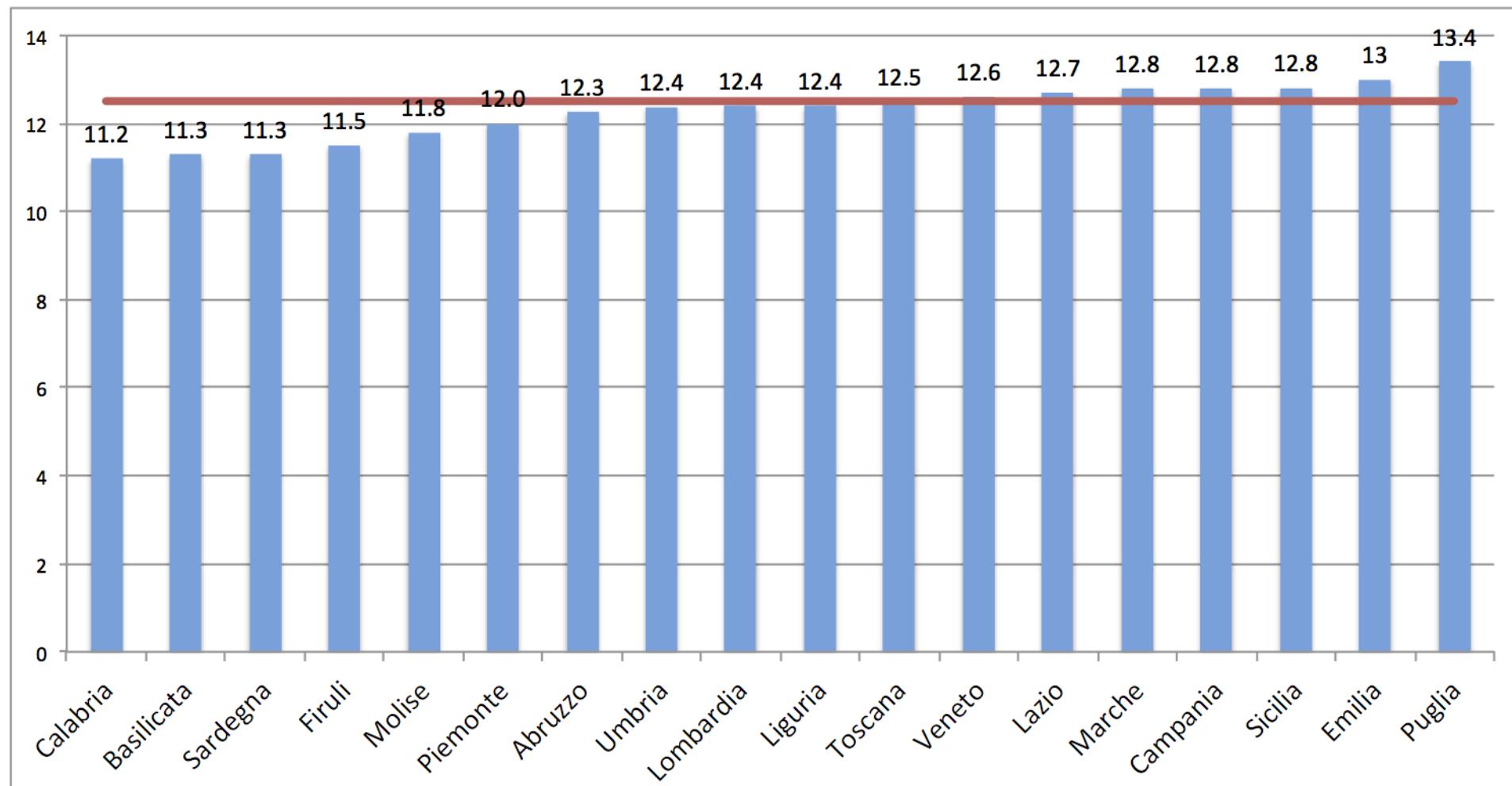
Clarity:
missing axis +
angled labels

Proportionality:

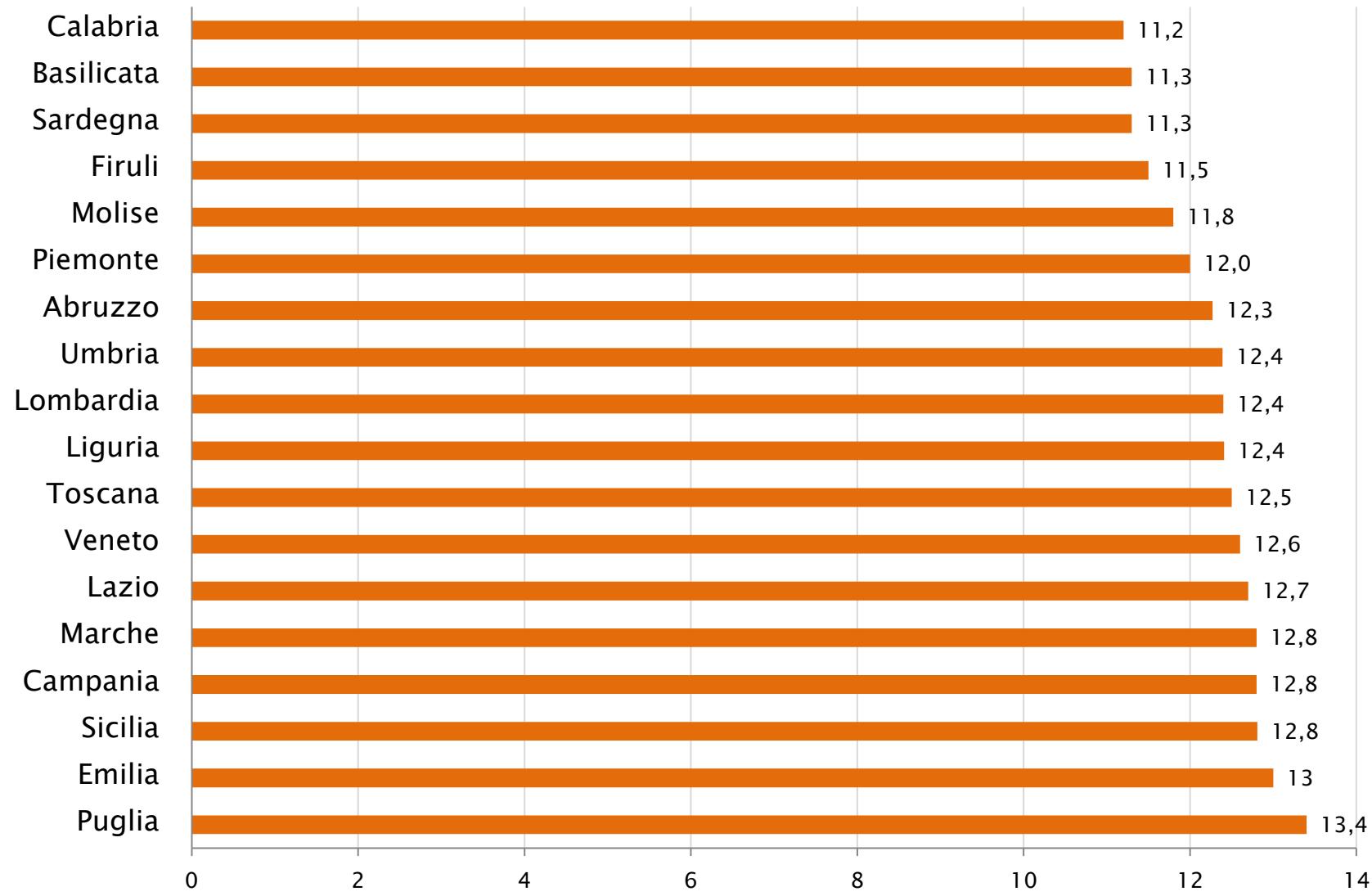
$$LF = \frac{13.4/11.2}{11.4/4.2} = 2.26$$



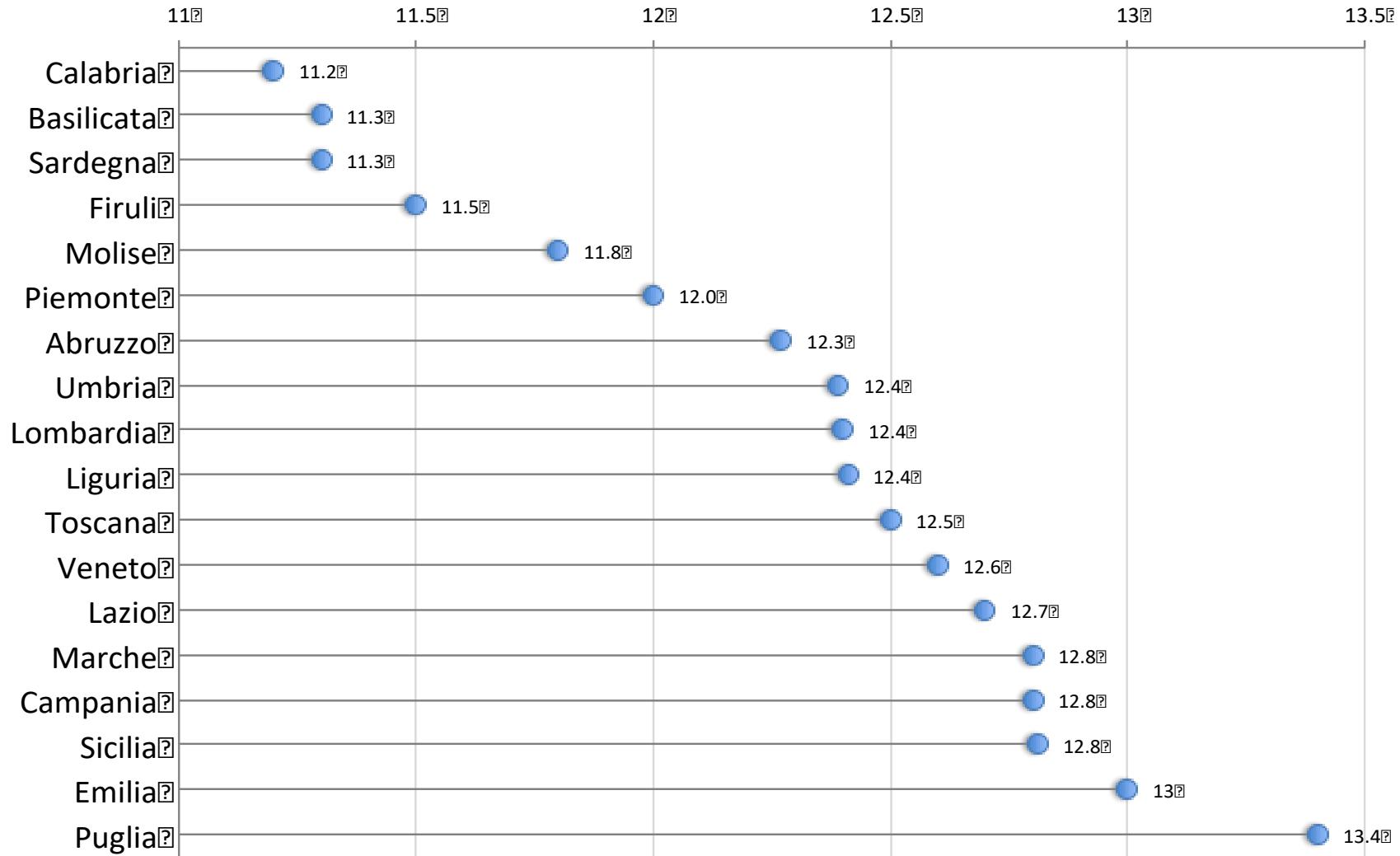
Bar are zero based



Horizontal let longer labels

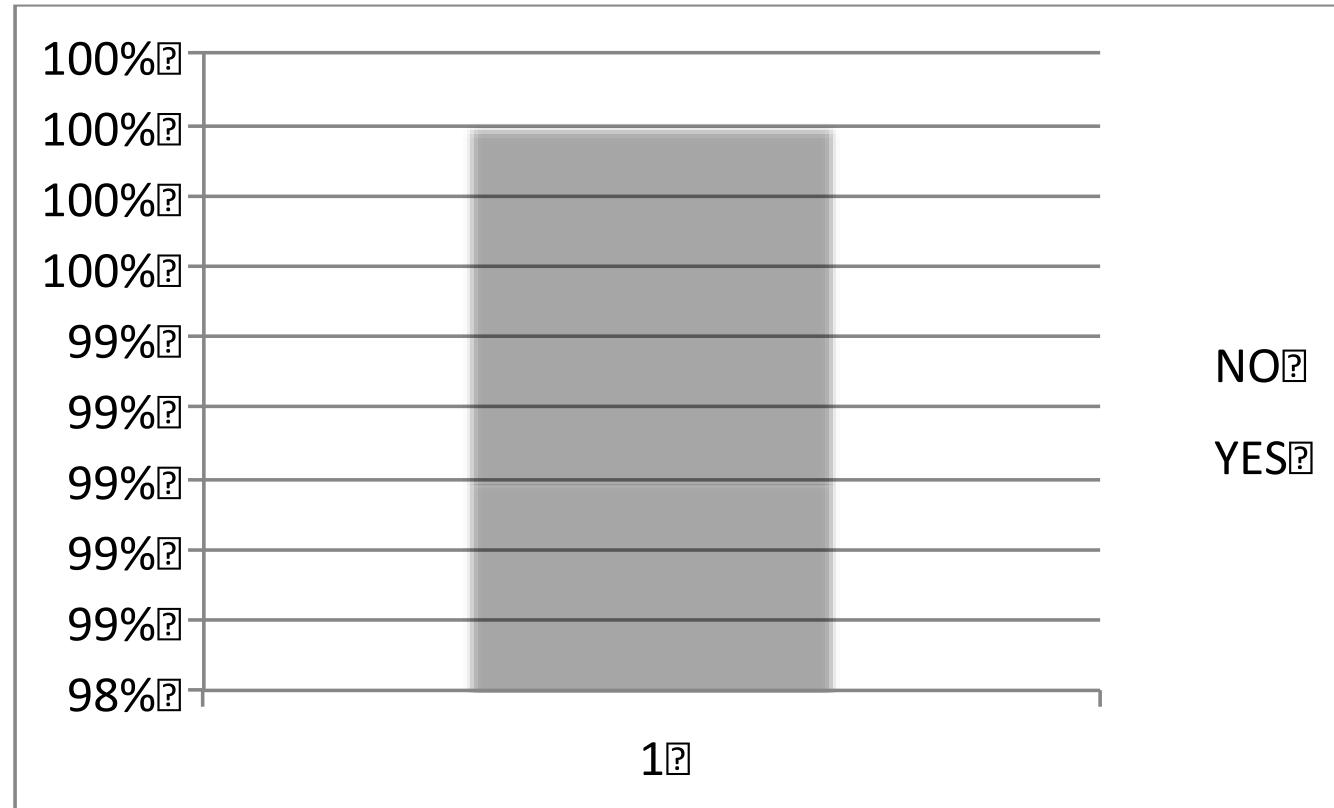
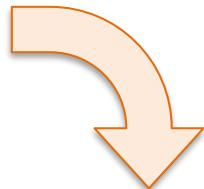


Dot plot (Lollipop)



Beware MS-Excel Default

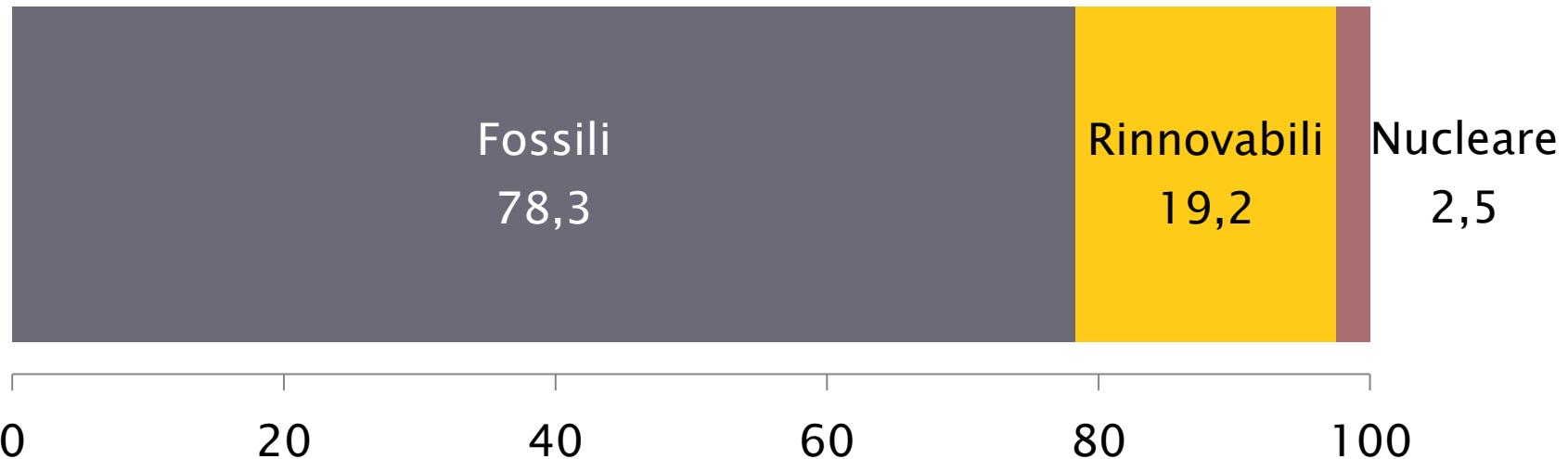
	A	B
1	YES	99%
2	NO	1%



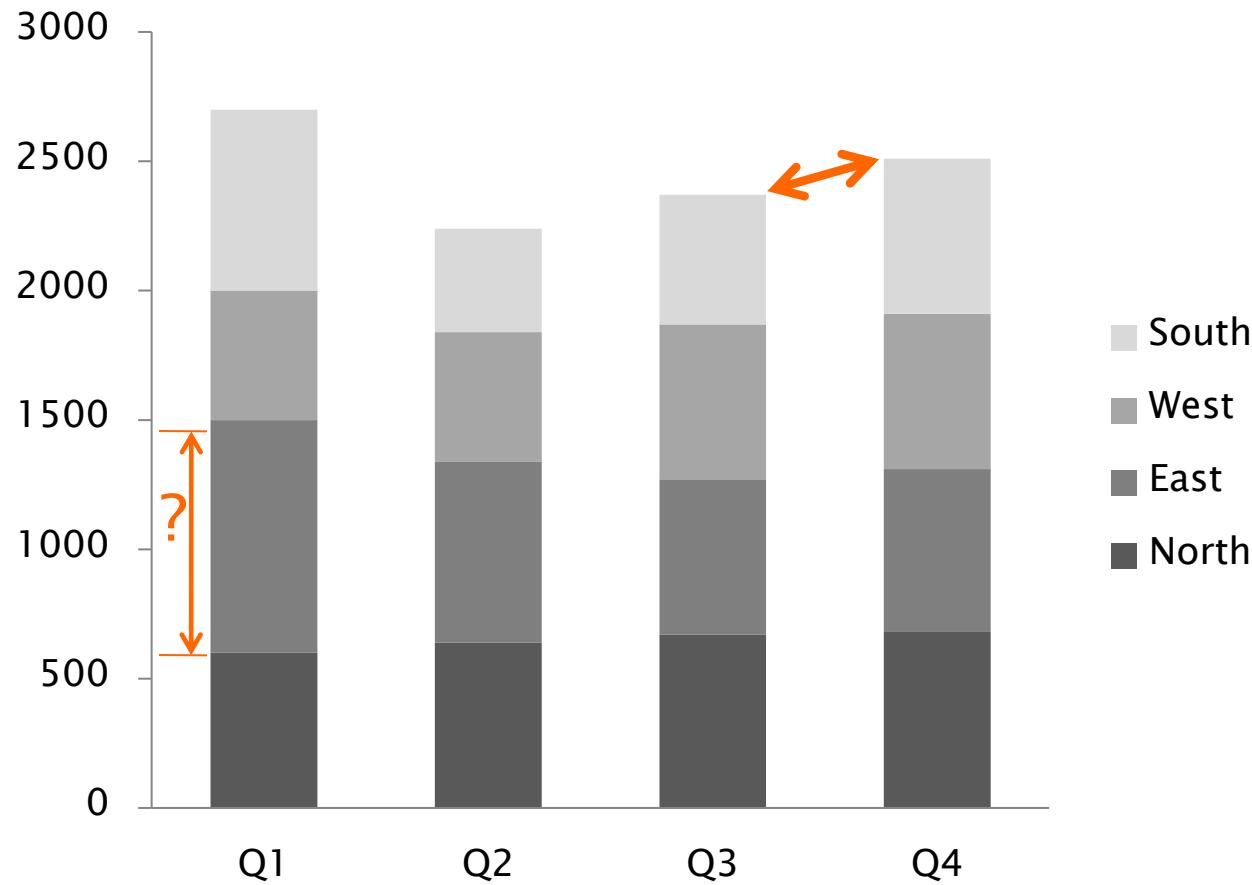
Bars Guidelines

- Use horizontal bars when
 - ◆ A descending order ranking
 - ◆ Categorical label don't fit
- Proximity
 - ◆ Use a 1:1 bar:spacing ratio $\pm 50\%$
 - ◆ No spacing between bars that are not labeled on the axis (legend categories)
 - ◆ No overlapping bars

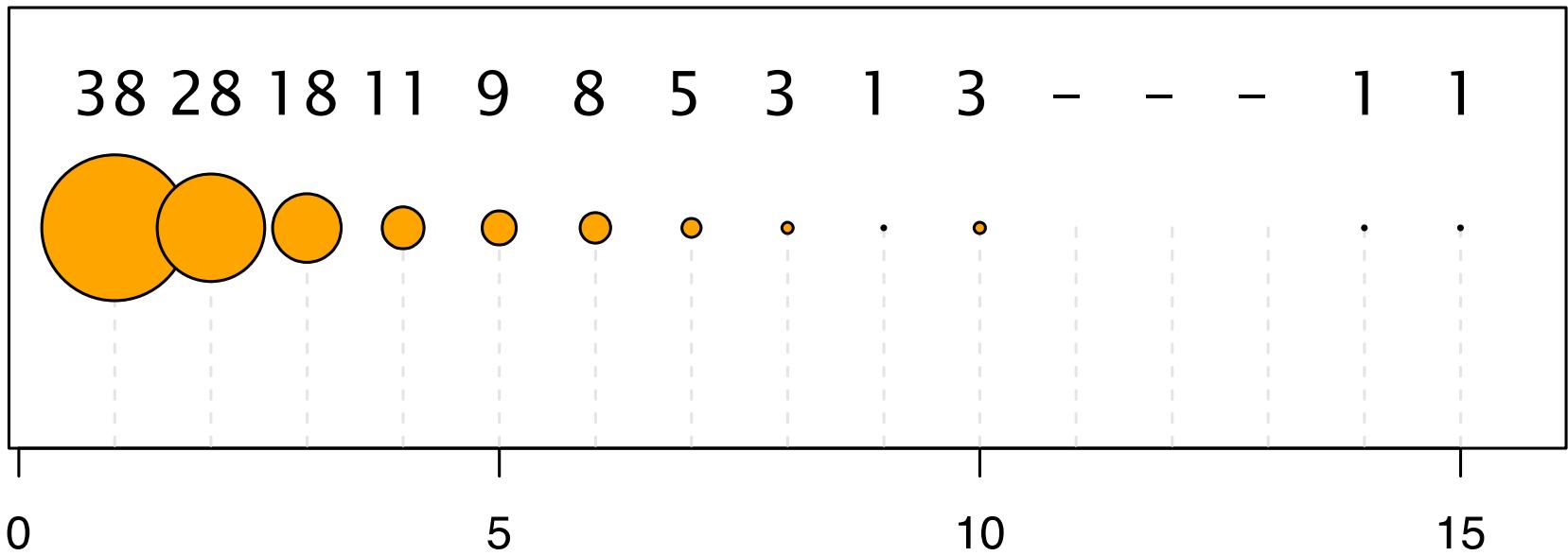
Length - Stacked Bars



Nonaligned bars – Stacked



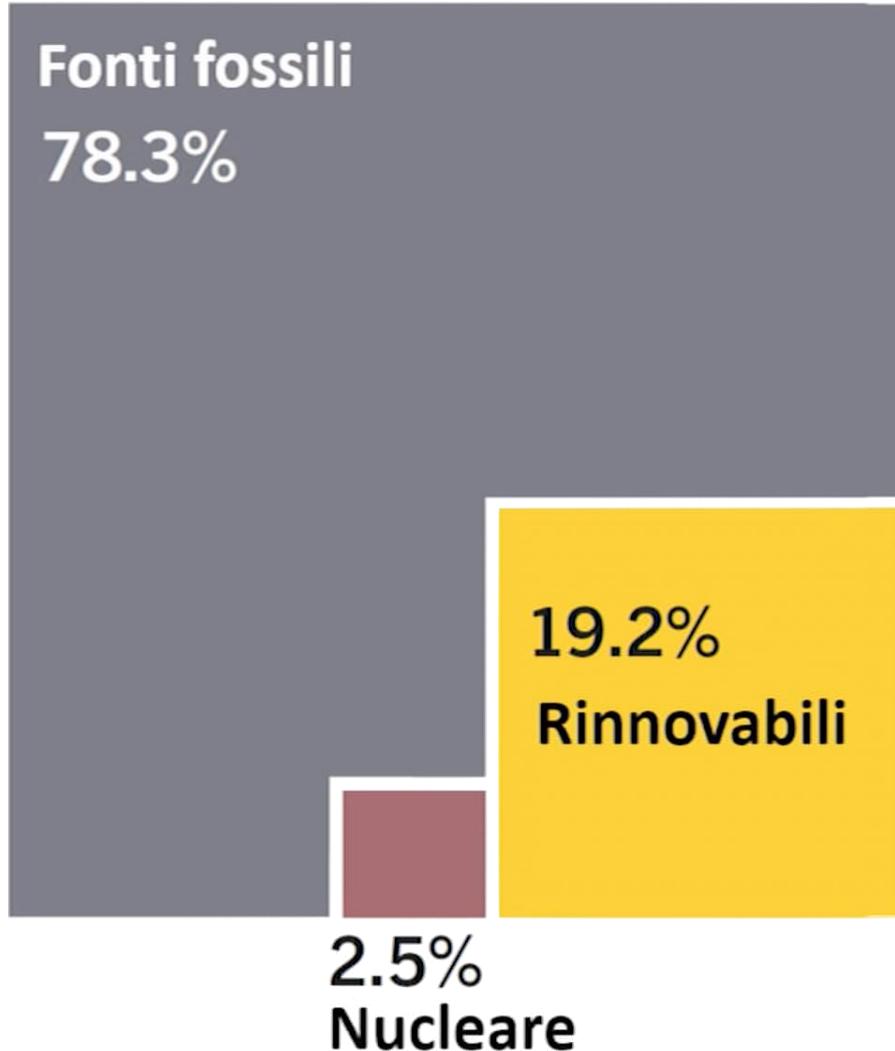
Shape area – Bubbles



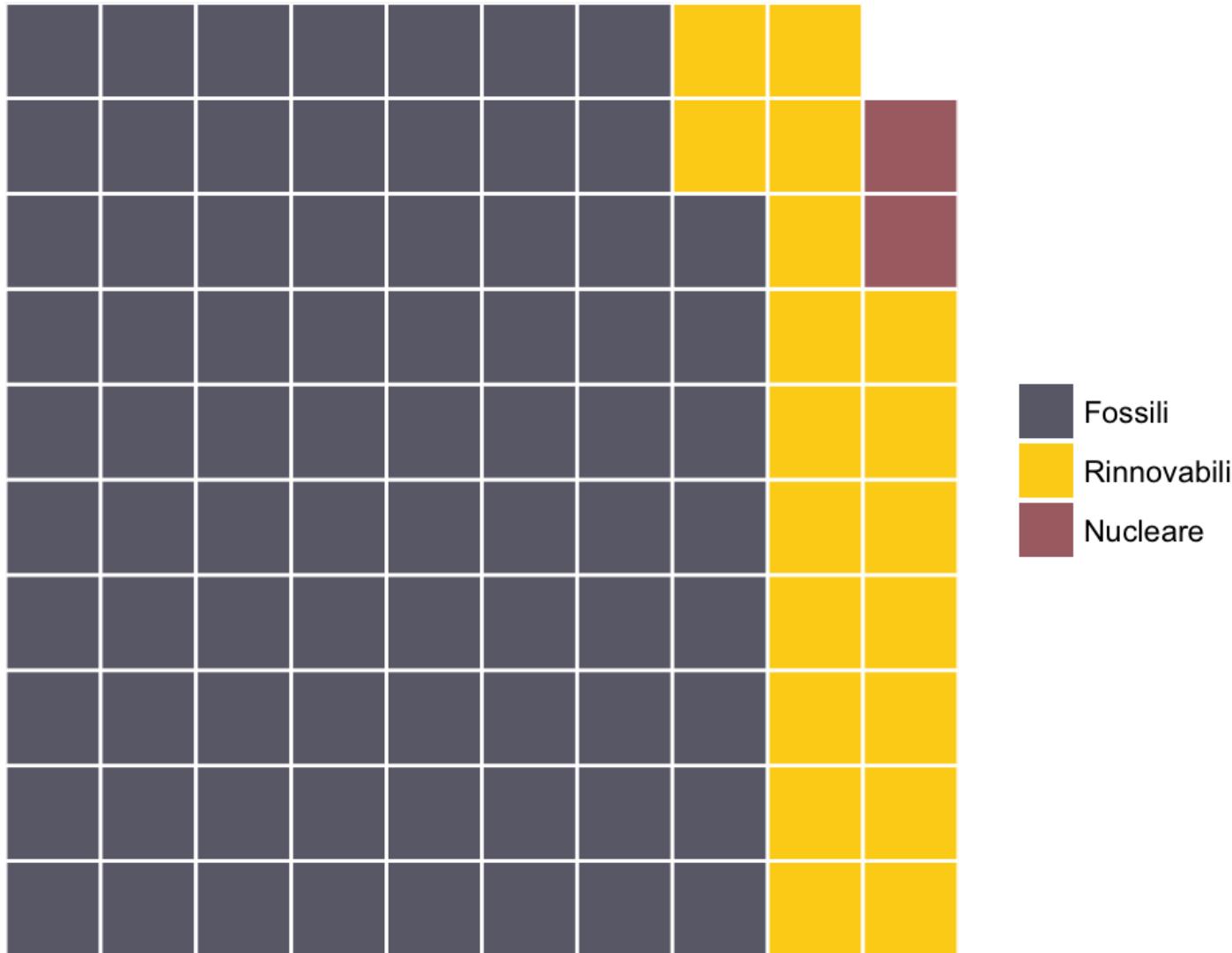
Area – Treemap



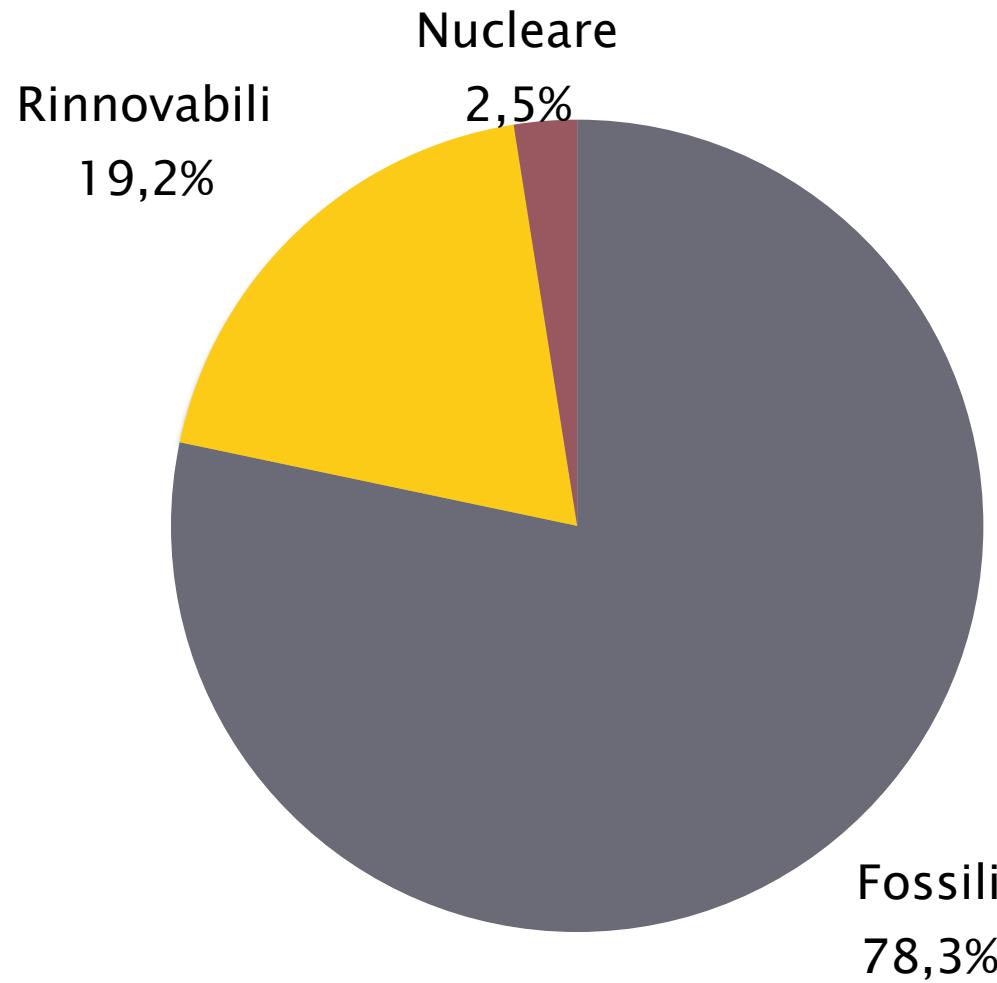
Area – Treemap



Area + Count – Waffle / Grid



Area + Angle – Pie Chart



Count – Isotype

- Isotype
 - ◆ International System Of Typographic Picture Education
- Marie and Otto Neurath
 - ◆ Vienna, 1936

Literacy in England and Wales

Among 10 men

Illiterates

1841



1871



1901



1931



Among 10 women

1841



1871



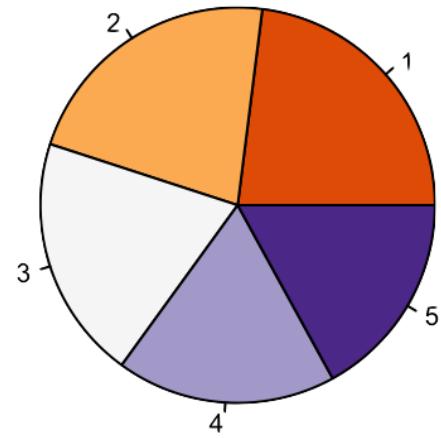
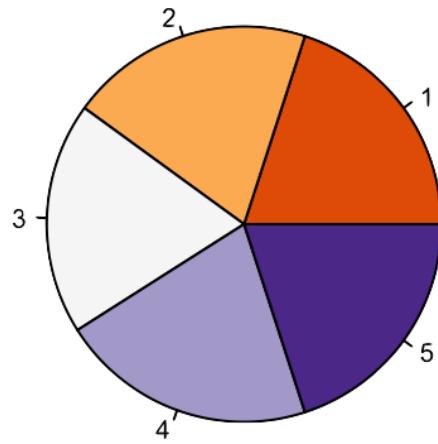
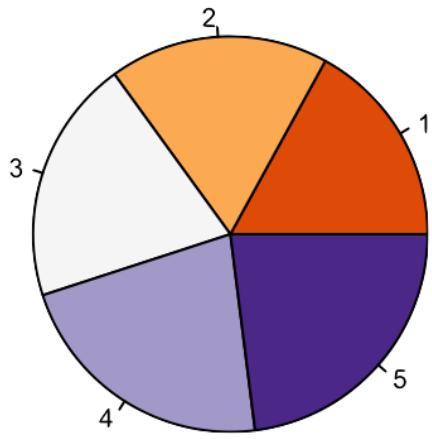
1901



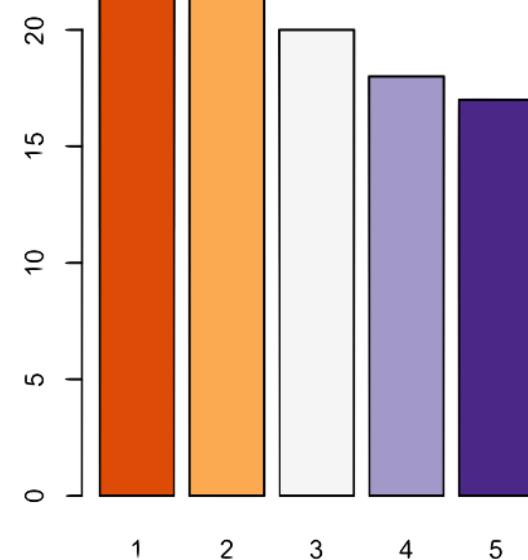
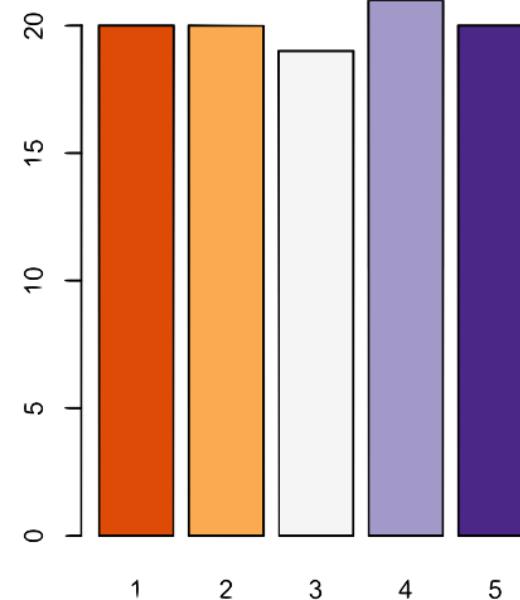
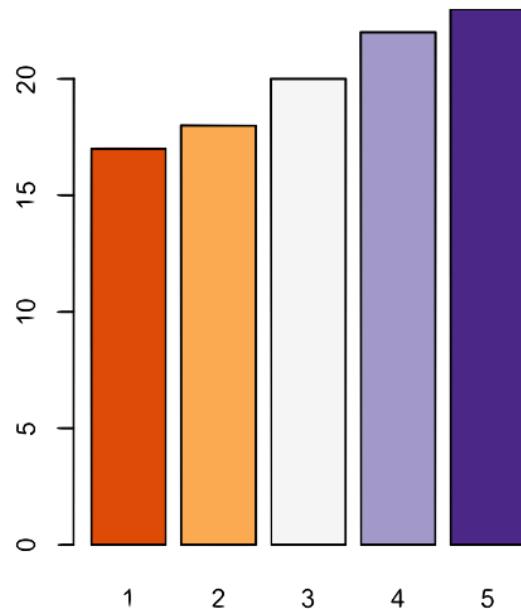
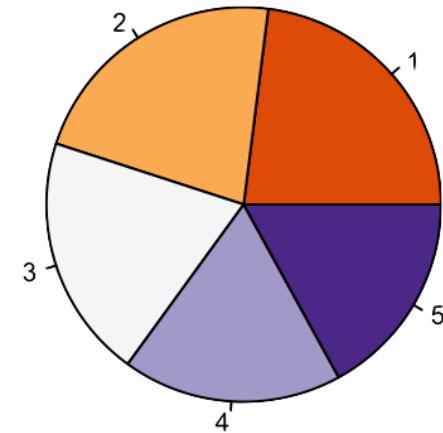
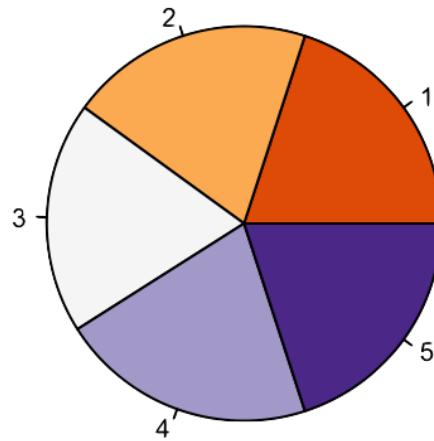
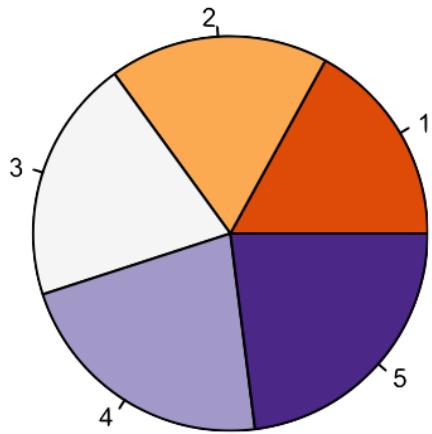
1931



Pies



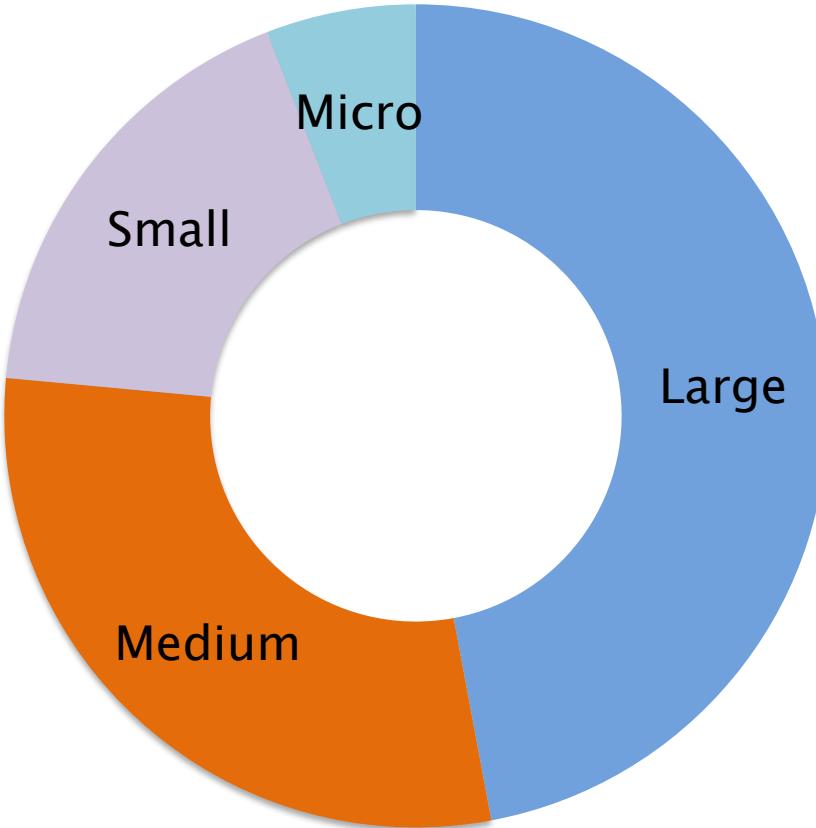
Pies vs. Bars



Pie Charts: guidelines

- Have serious limitations
 - ◆ To represent part–whole relationship
 - ◆ Only with a small number of categories
 - Up to four
 - Avoid rainbow pie
 - ◆ When proportions are distinct enough
- Remember to ease reading
 - ◆ Labels placed close to slices
 - ◆ Labels include values (percentages)

Area/Angle/Length – Donut



Categorical encoding attributes

- Encoding of categorical levels

- ◆ Position (along an axis)

- ◆ Size

- ◆ Color

- Intensity

- Saturation

- Hue

- ◆ Shape

- ◆ Fill pattern

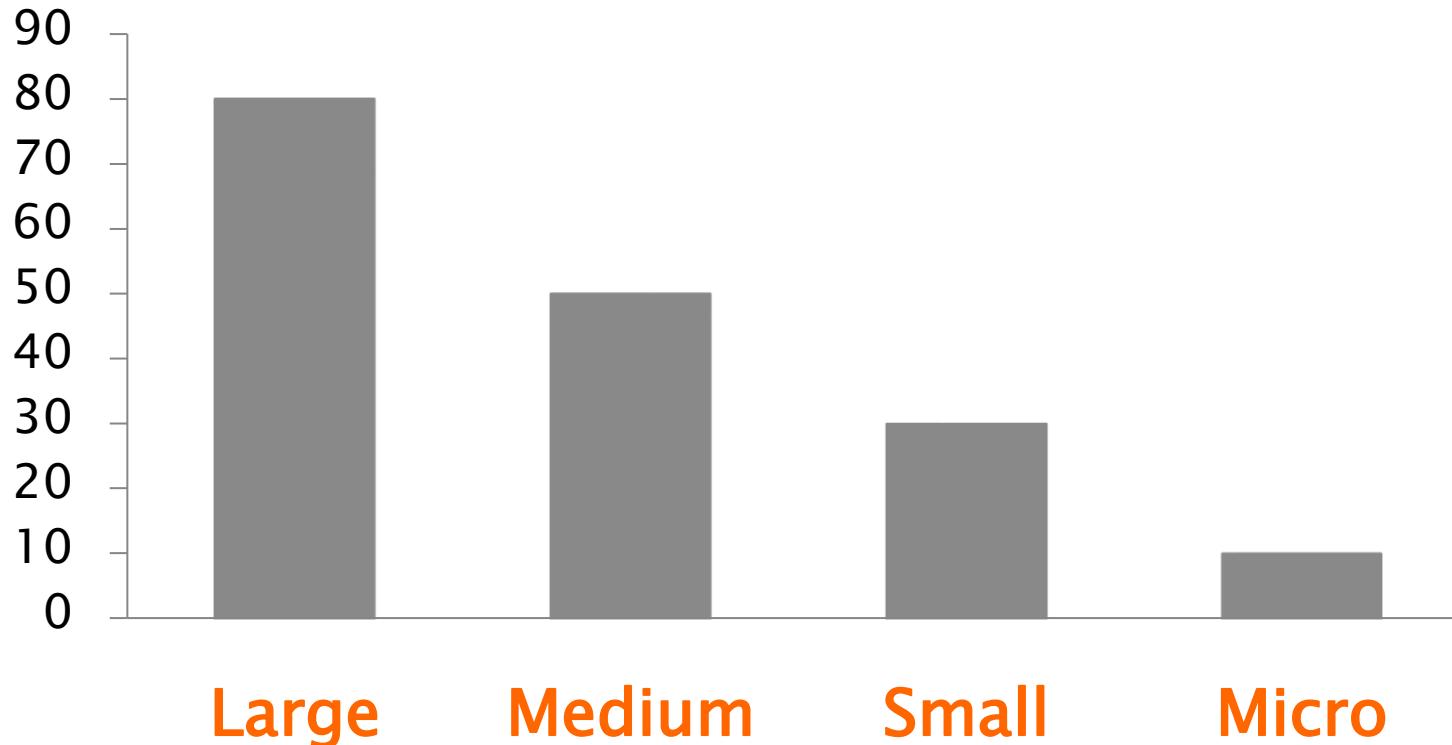
- ◆ Line style



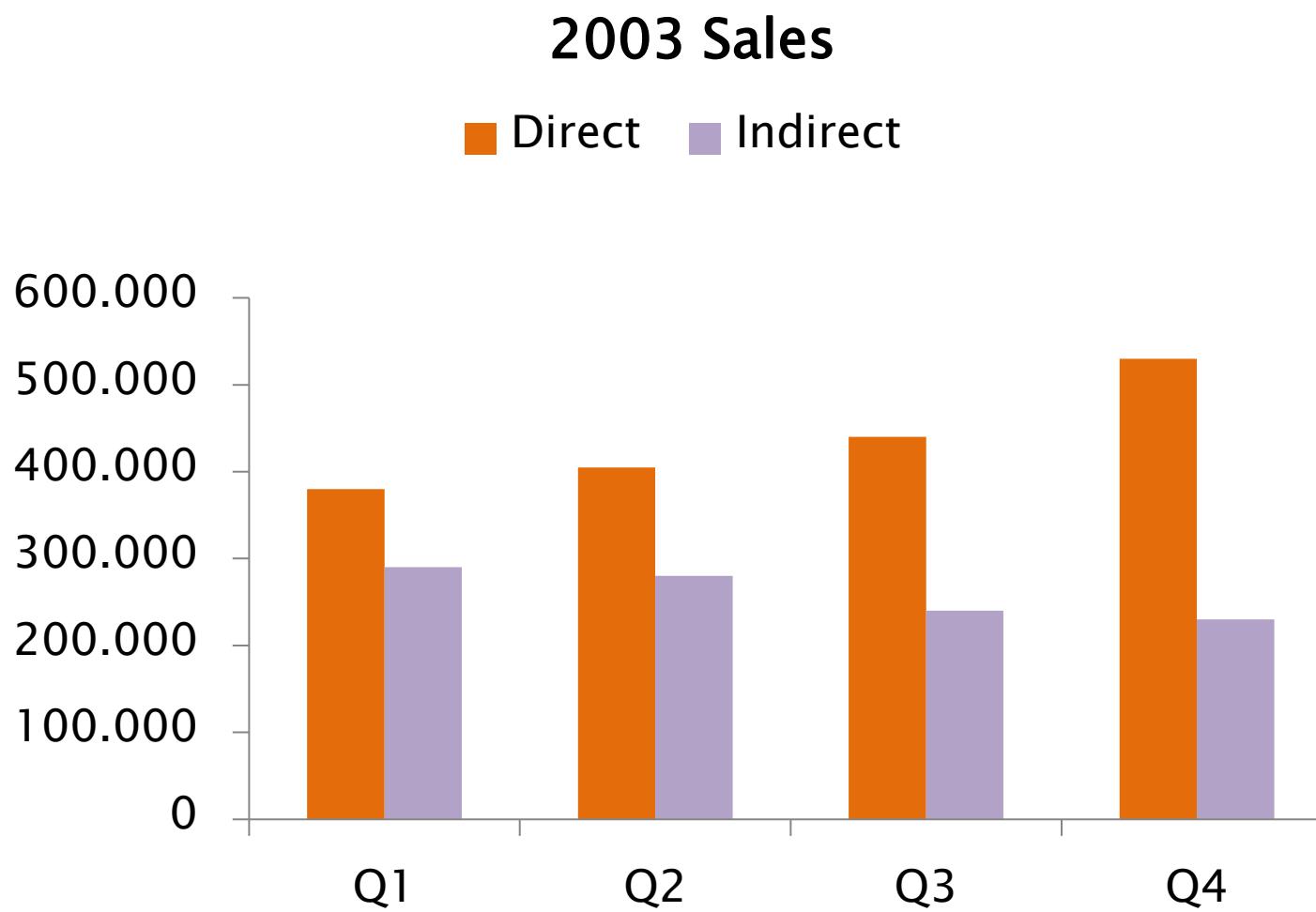
Ordinal

Position

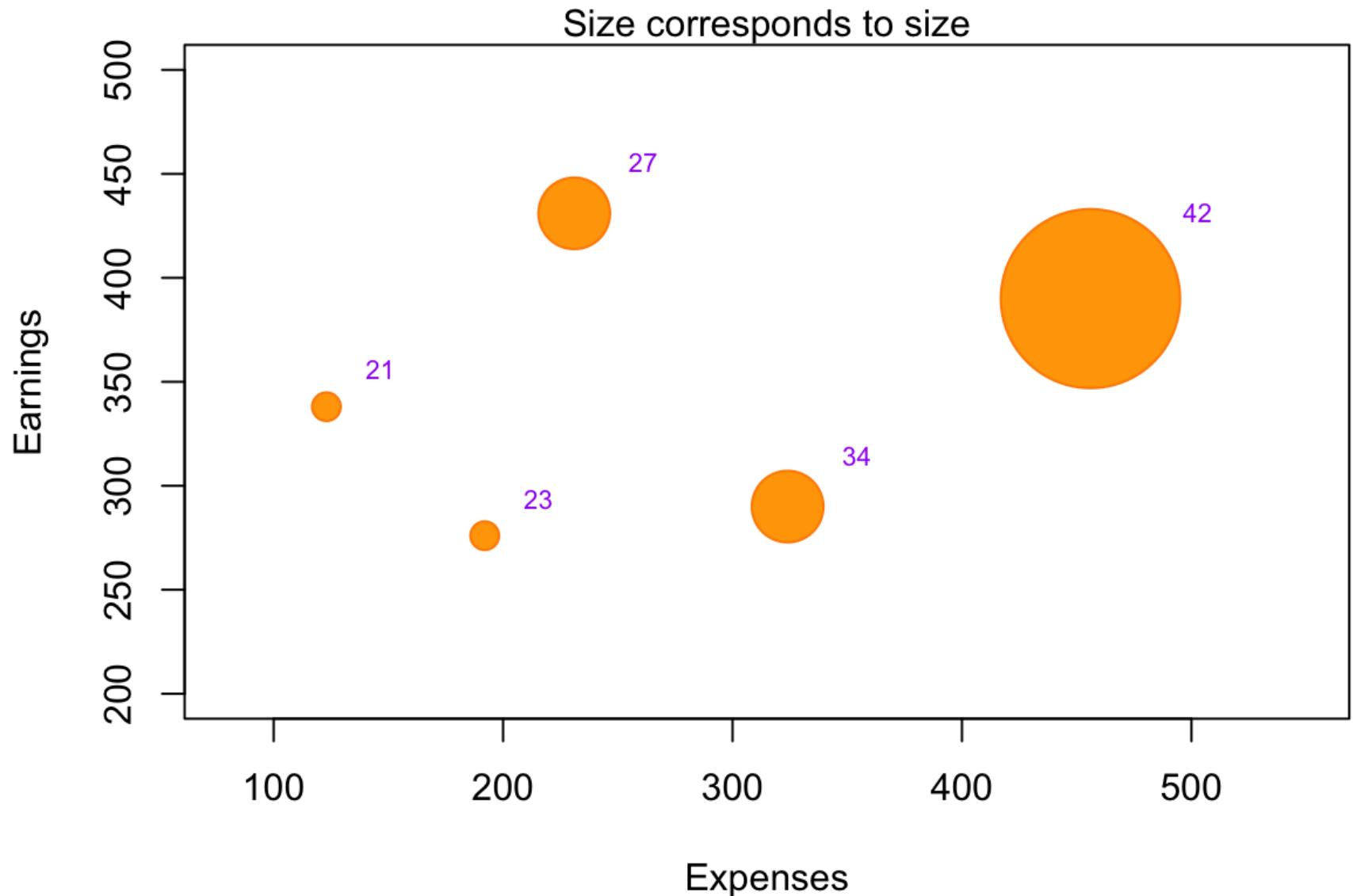
Number of
companies



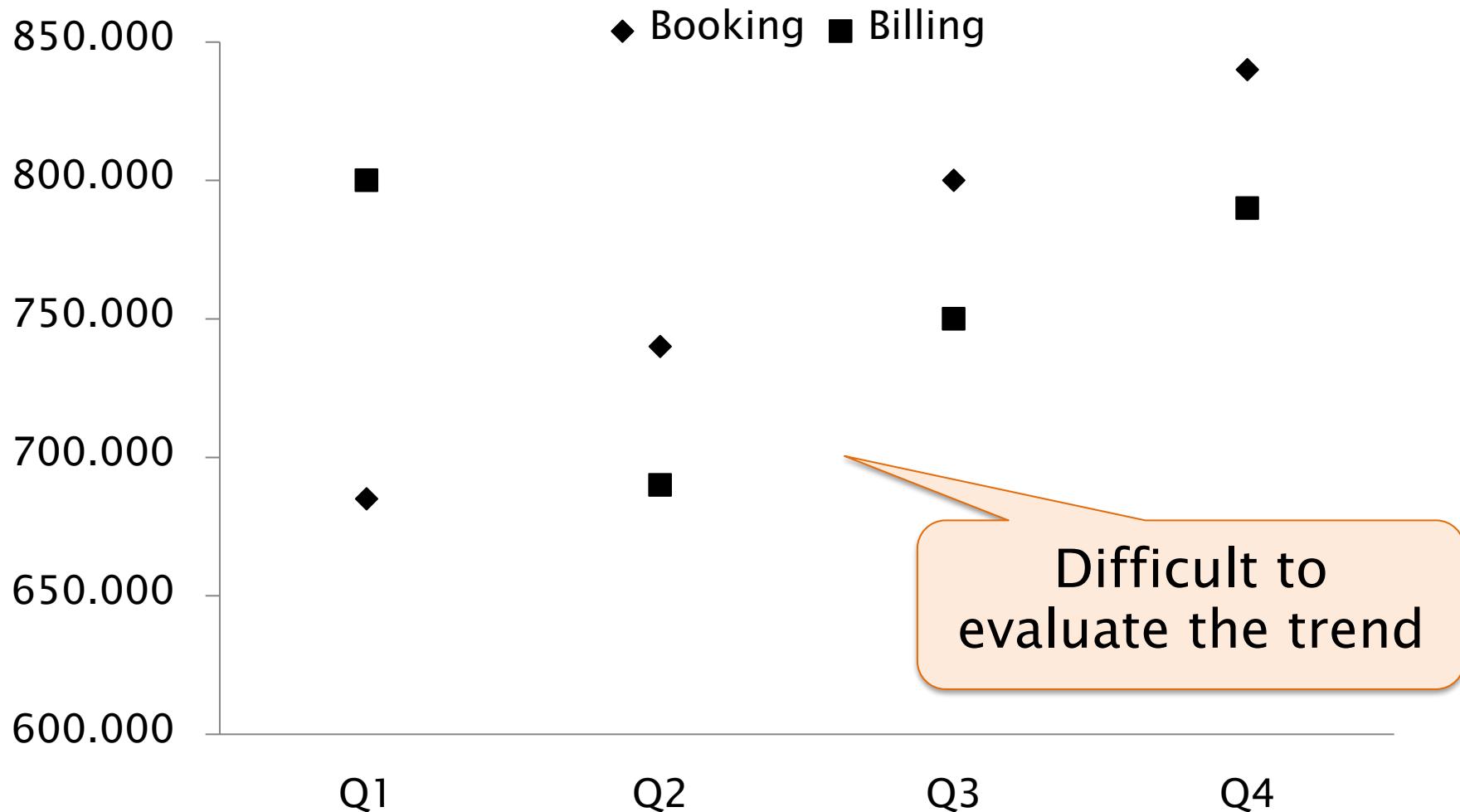
Position + Color (hue)



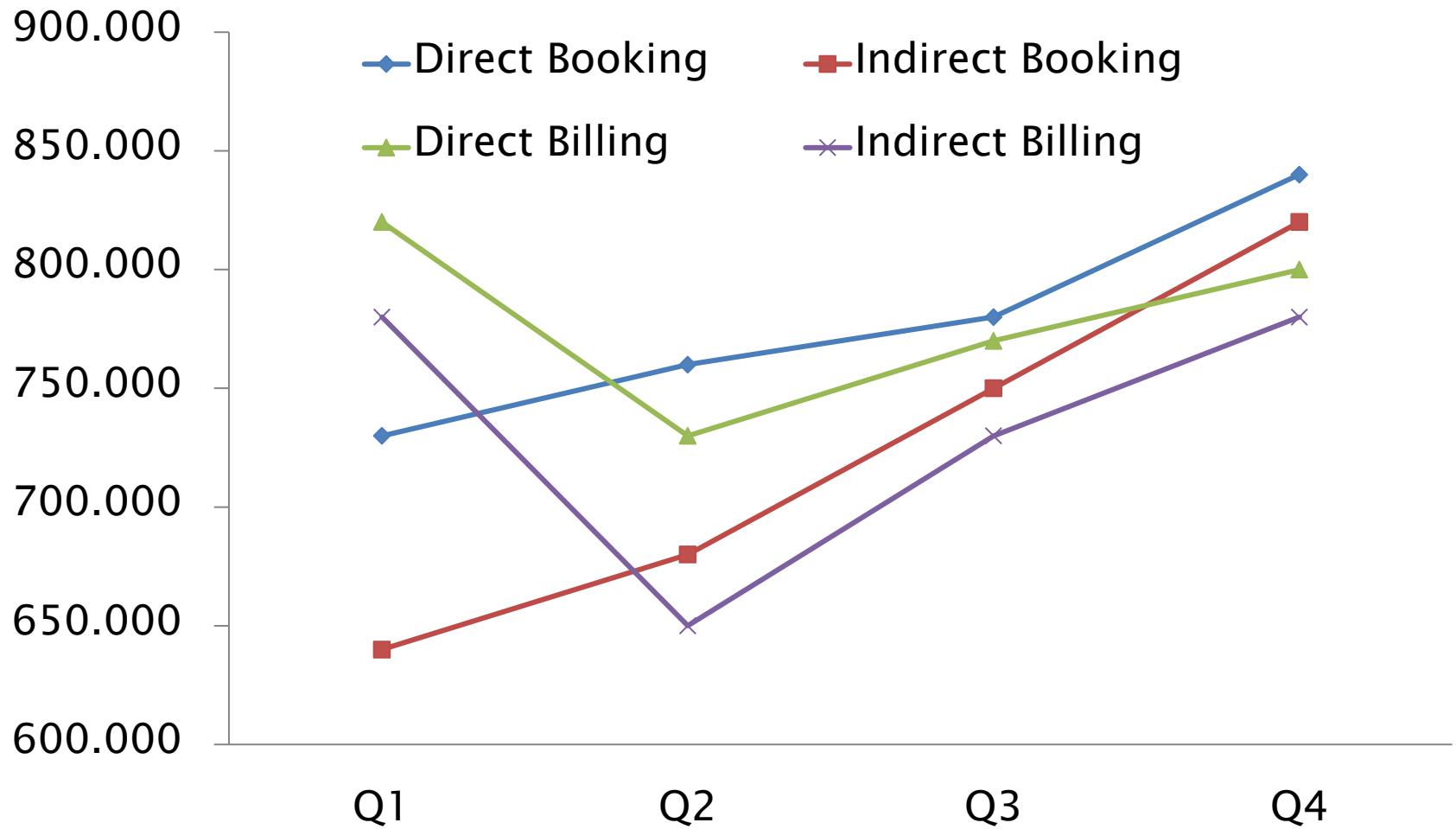
Size



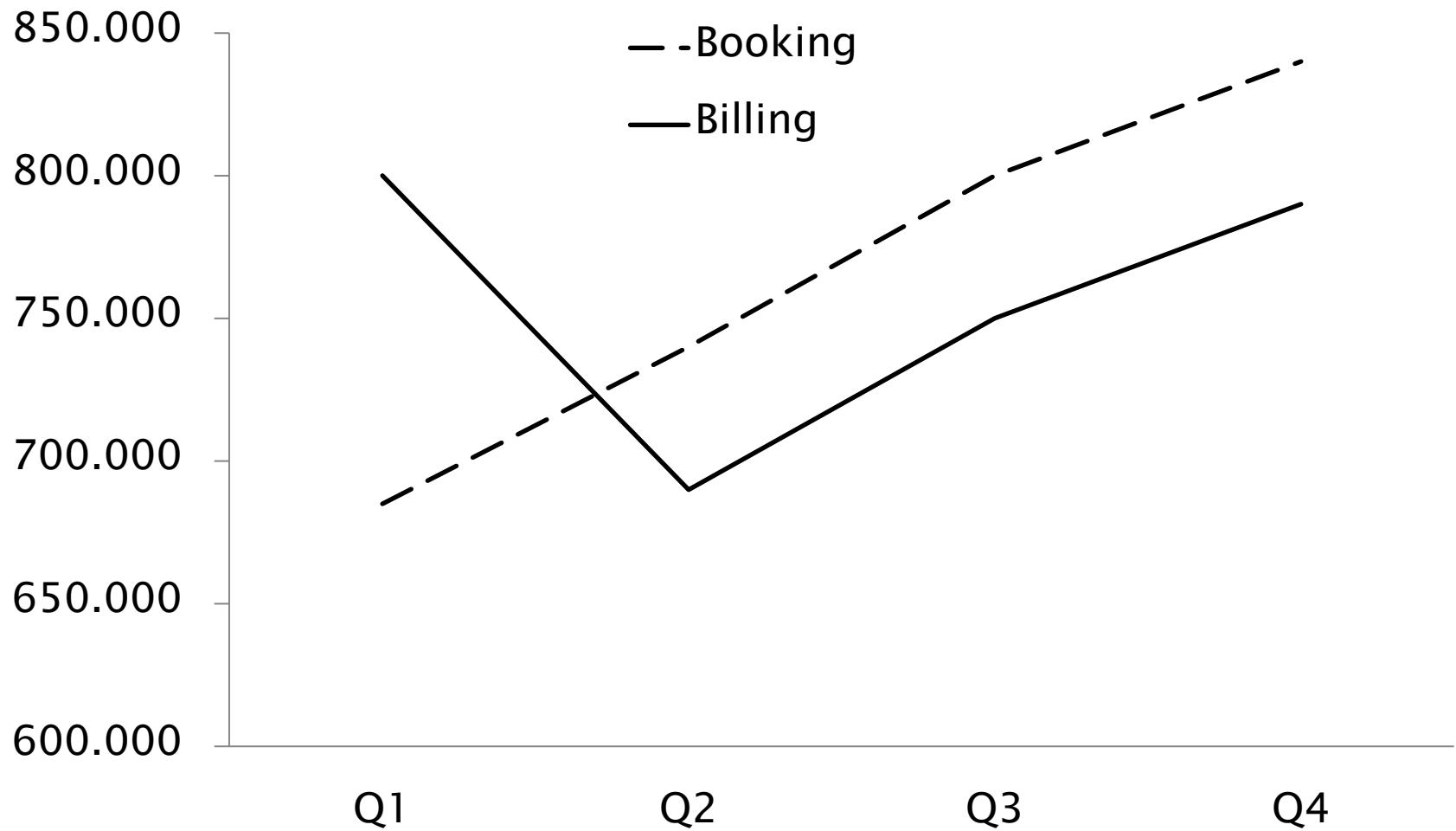
Point shape



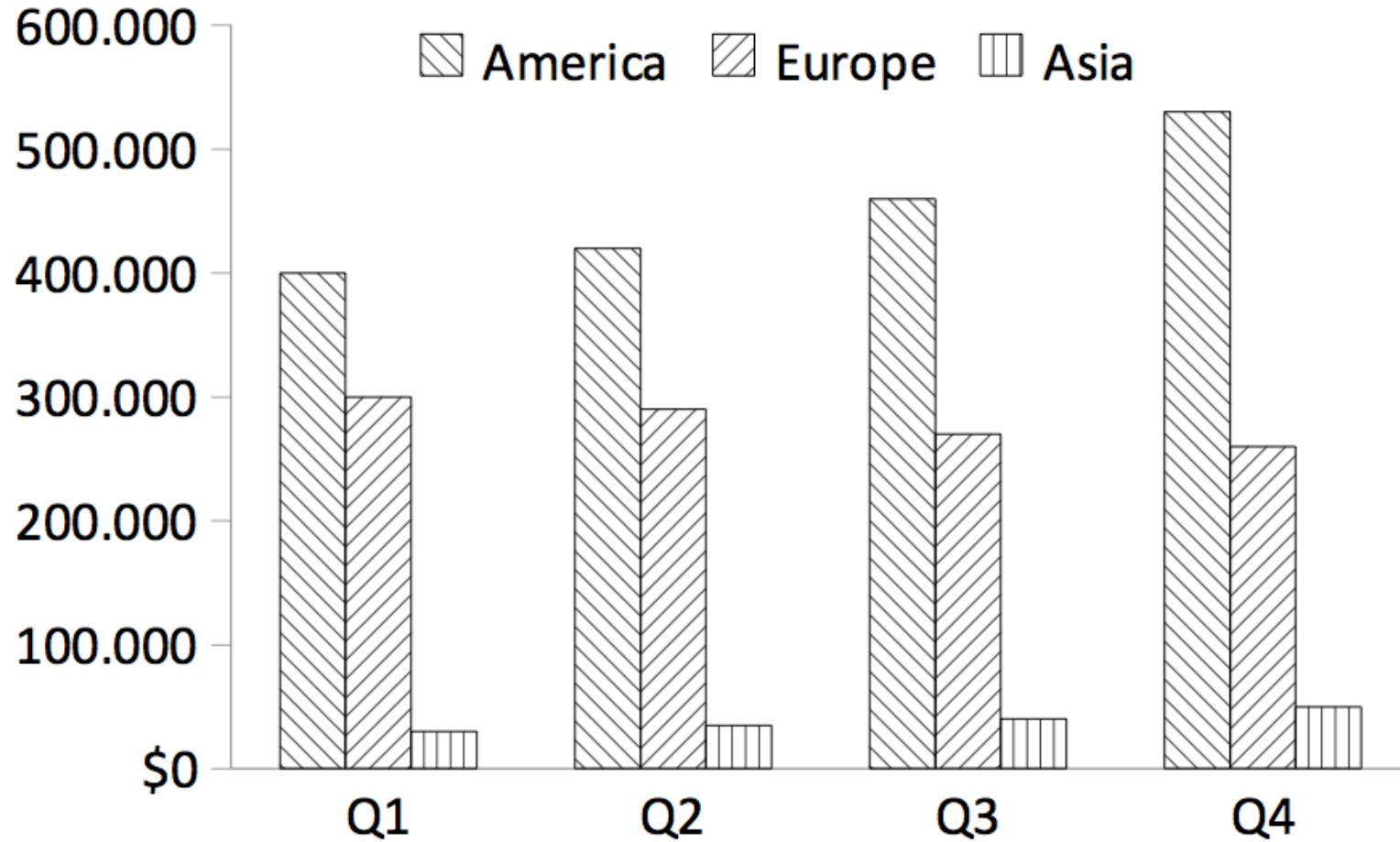
Point shape + Color



Line style



Fill Texture

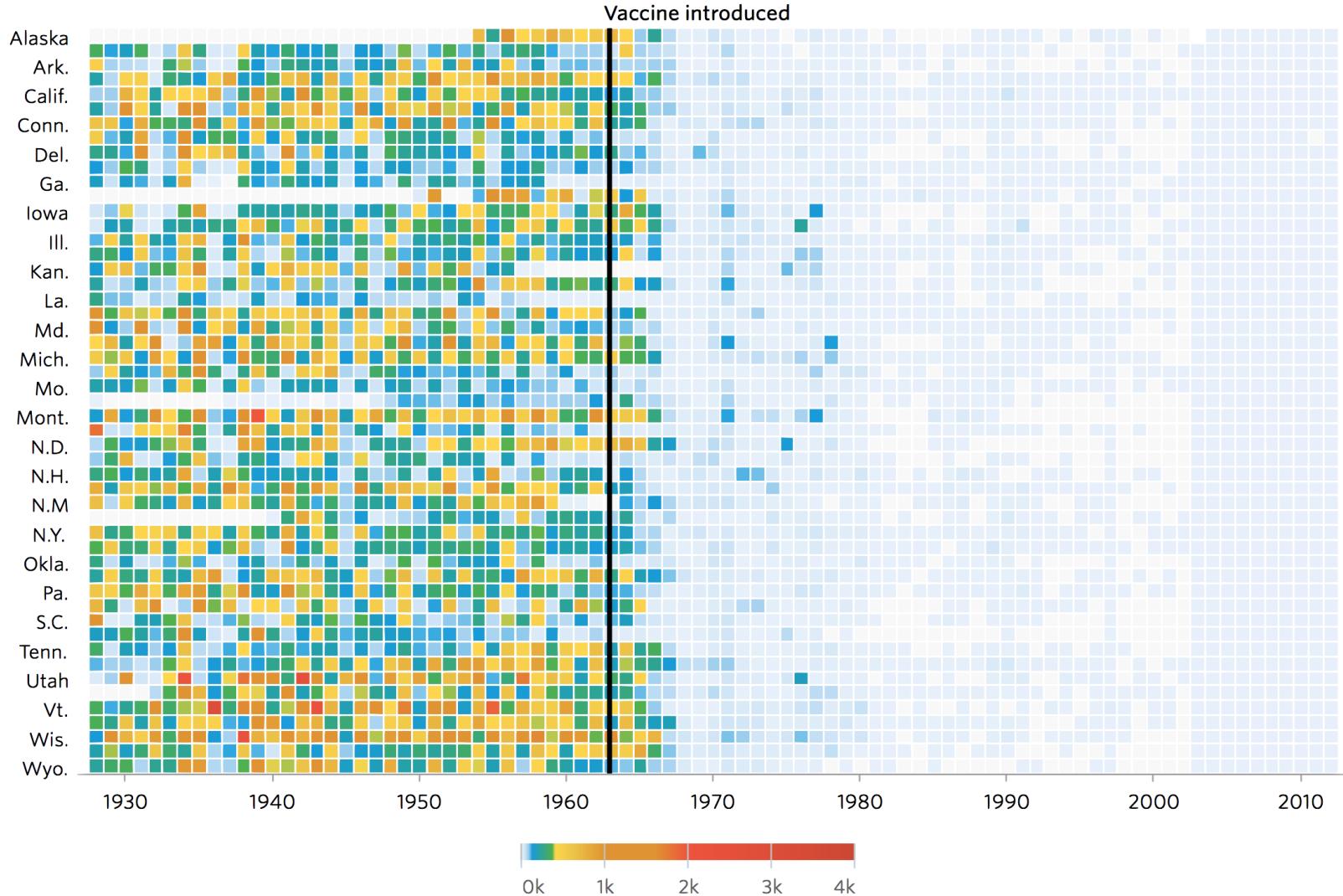


Discretization / Quantization

- A data transformation that maps a quantitative measure into an ordinal one
 - ◆ Based on the definition of intervals
- Discretized measures can be encoded using an ordinal-friendly visual attribute
 - ◆ Size
 - ◆ Color
- Warning: details are lost in the process

Heatmaps

Measles

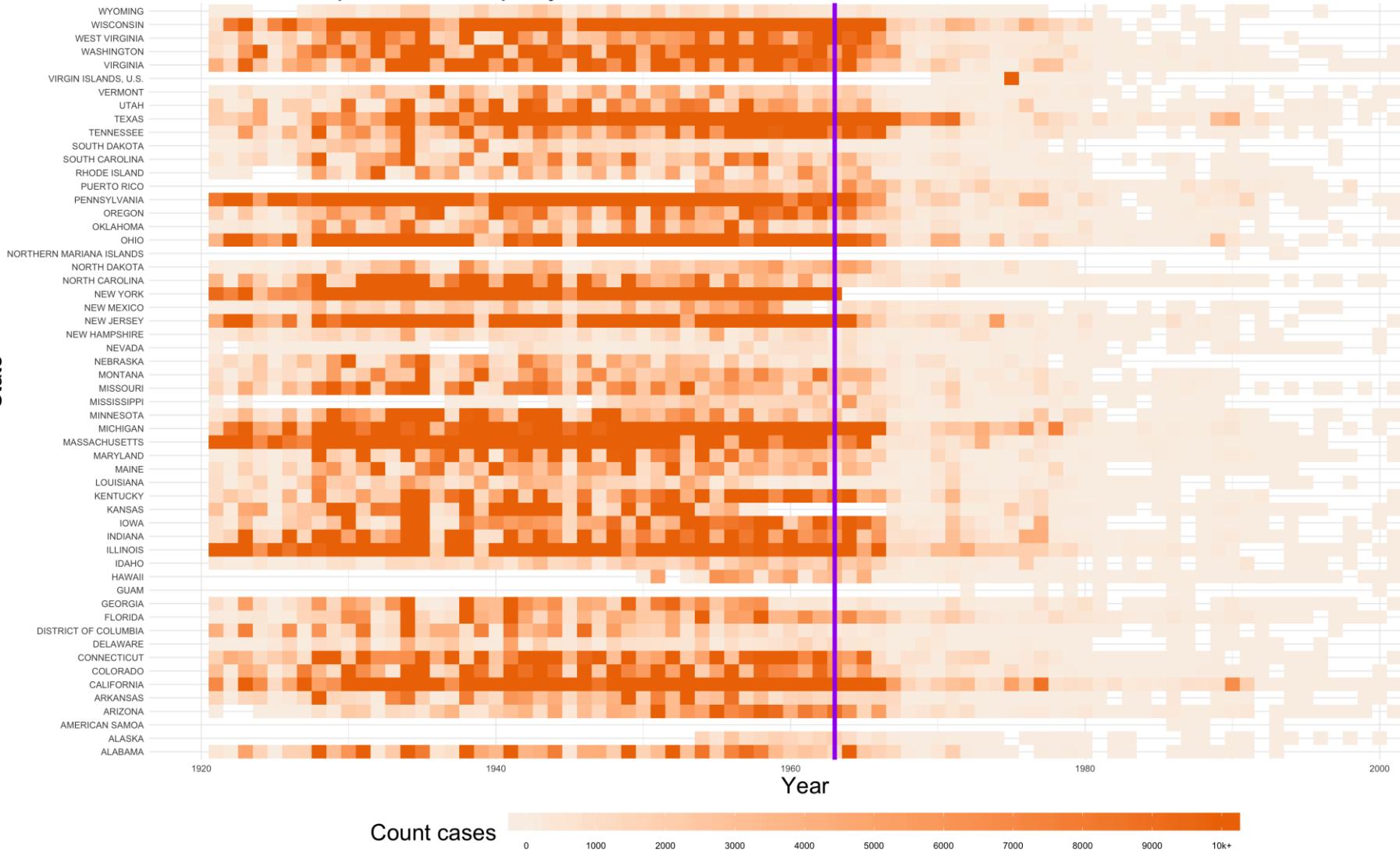


Heatmaps

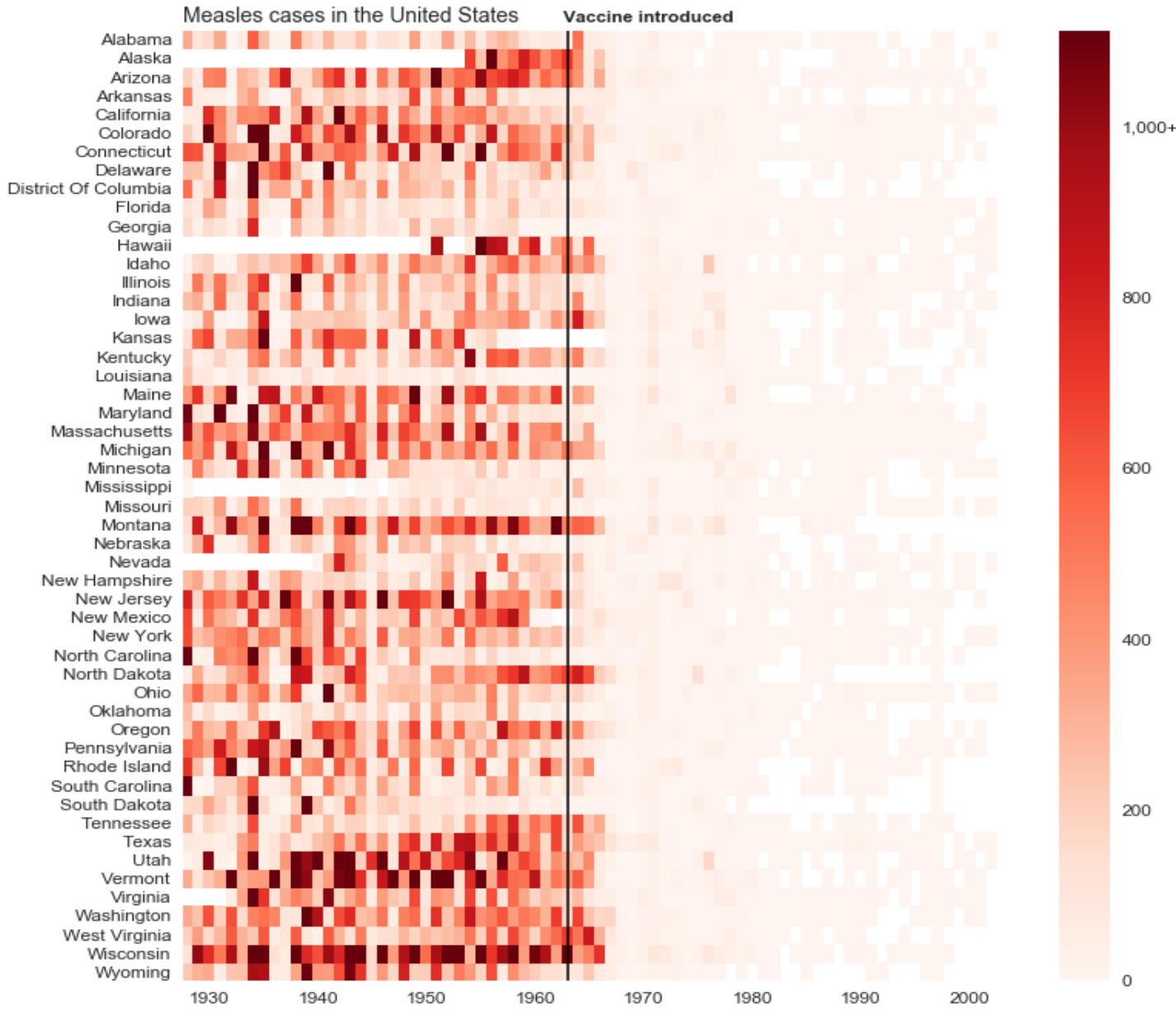
- Hues have no unique order semantics
 - ◆ Only intensity has one
- Rainbow palette have serious problems for color blinds
 - ◆ Roughly 5% of the population

Heatmaps

Measles cases per US State per year



Heatmap



Data source: Project TYCHO (tycho.pitt.edu) | Author: Randy Olson (randalolson.com / @randal_olson)

<http://www.randalolson.com/2016/03/04/revisiting-the-vaccine-visualizations/>

SUPPORT ELEMENTS

Support elements

- Axes
 - ◆ Ticks
- Graph area
 - ◆ Grids
- Labels
- Legends
- References
- Trellies

Axes

- Allow positioning of elements
 - ◆ Points
 - ◆ Extremes of bars and lines
- Labeled
 - ◆ What is the measure?
- Number of axis should be 2
 - ◆ 1 is fine for bars
 - continuity gestalt principle

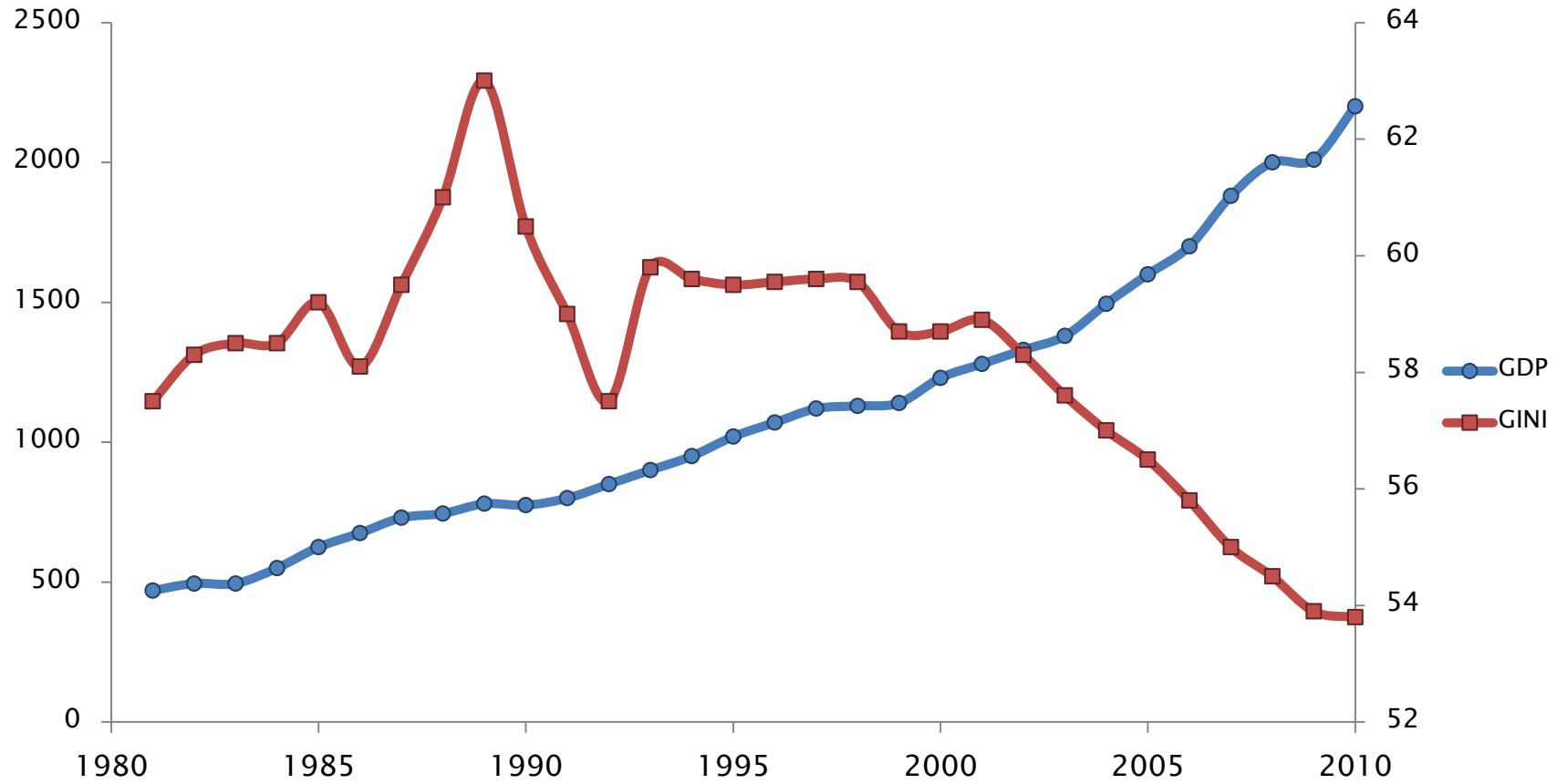
Tick marks

- Must not obscure data objects
- Outside the data region
- Avoid for categorical scales
- Balanced number
 - ◆ Too many clutter the graph
 - ◆ Too few make difficult to discern reference for data objects
 - ◆ Intervals must be equally spaced

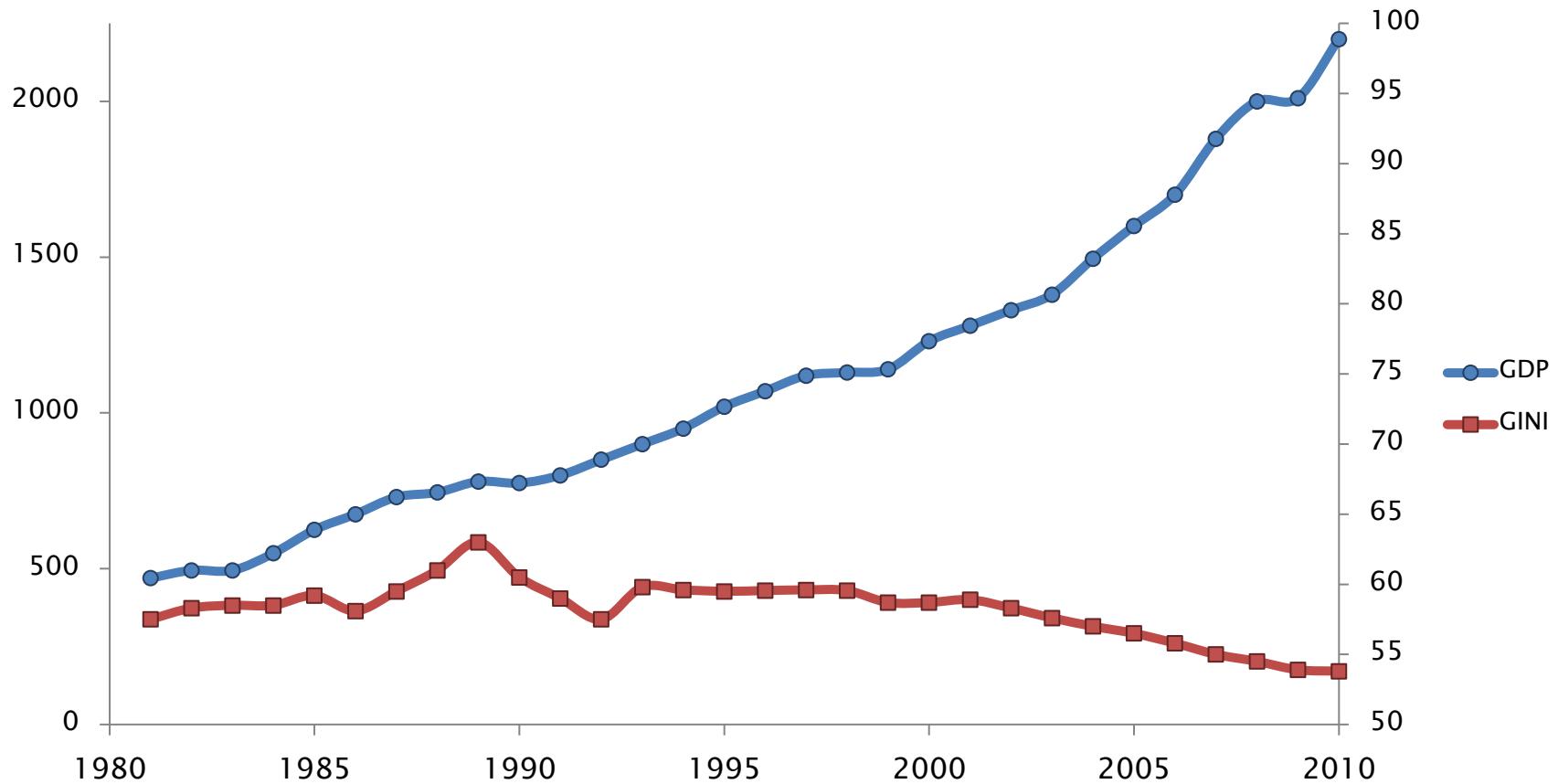
Multiple variables

- Correlation between 3+ variables
 - ◆ E.g. two measures in time series
- Multiple units of measure
 - ◆ Double quantitative (y) axis
 - ◆ Multiple graphs
 - ◆ One variable not encoded explicitly

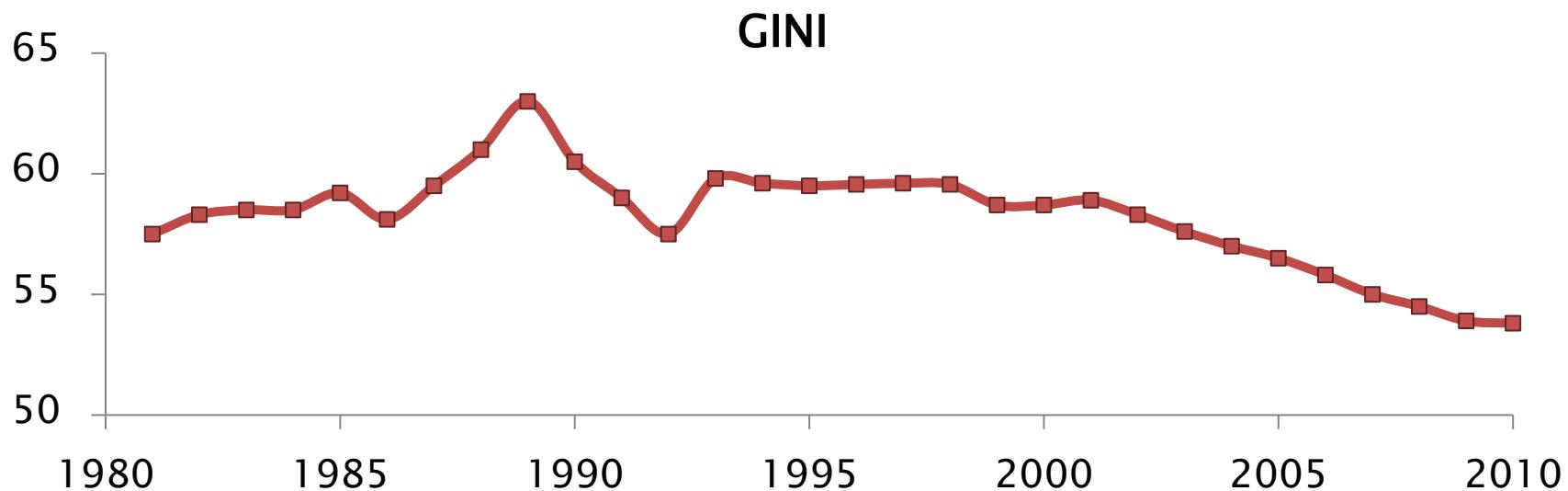
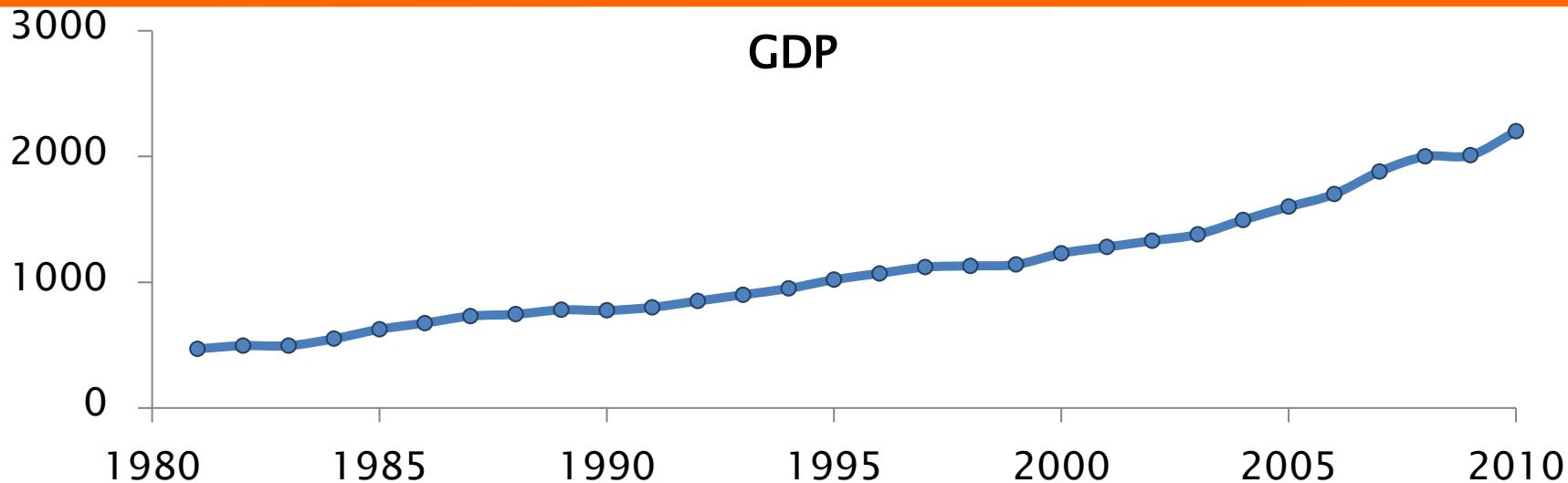
Double scale



Double scale



Multiple graphs



Small multiples

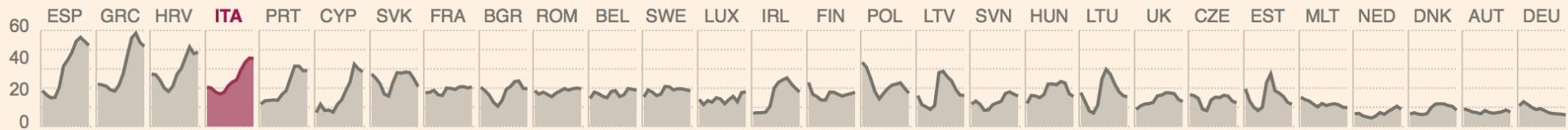
- A.k.a.
 - ◆ Trellis
 - ◆ Lattice
 - ◆ Grid
- Set of aligned graphs sharing (at least one) scale and axis
 - ◆ Enable ease of comparison among different measures

Small multiples

Total unemployment rate, 2004-2015 (%)



Youth unemployment rate, 2004-2015 (%)



Long-term unemployment rate, 2004-2014 (%)



FT EU unemployment tracker

<http://blogs.ft.com/ftdata/2015/04/17/eu-unemployment-tracker/>

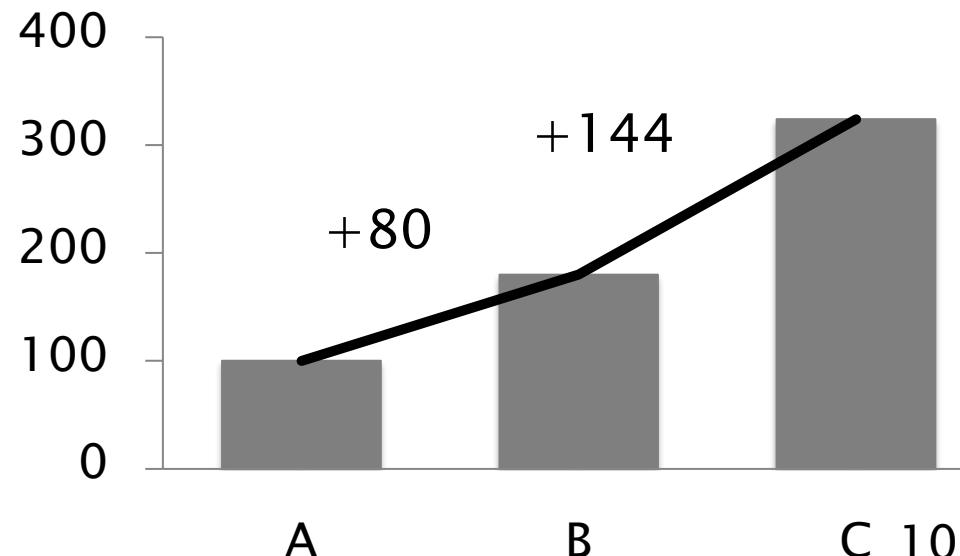
Trellis

- Sequence
 - ◆ Intrinsic order
 - ◆ Order of relevance
 - ◆ Order by some quantitative attribute
- Rules and grids
 - ◆ Use when spacing is not enough
 - ◆ Can direct the reader to scan graphs horizontally or vertically

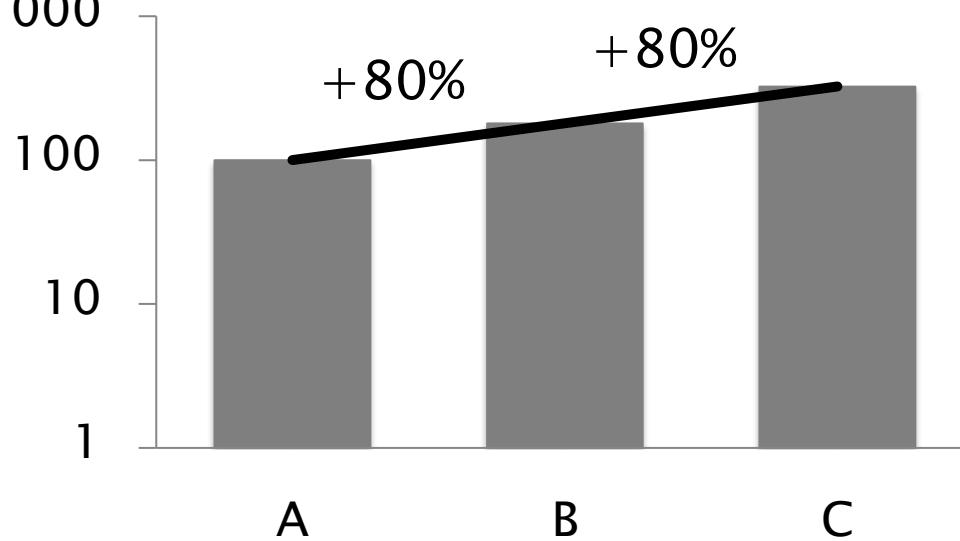
Log scale

- Reduce visual difference between quantitative data sets with significantly wide ranges
- Differences are proportional to percentages

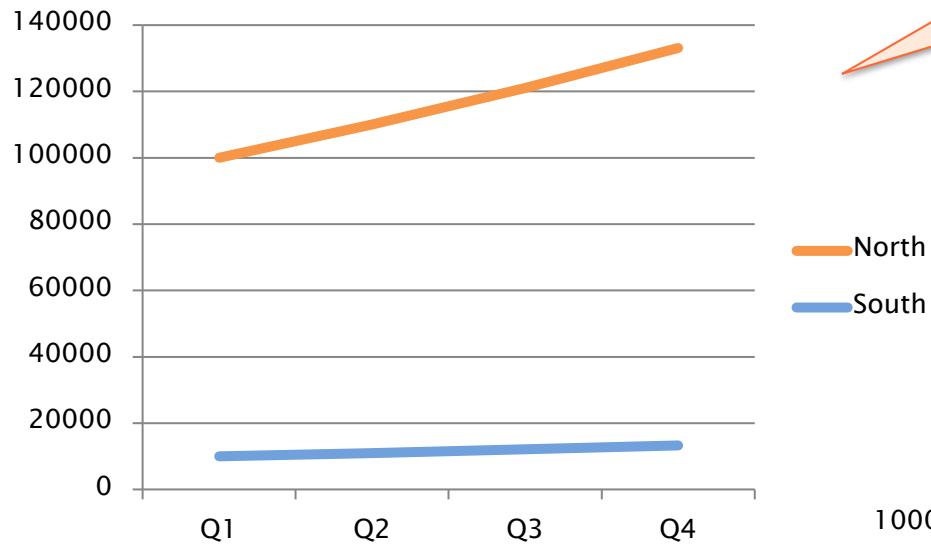
Log scale



C 1000

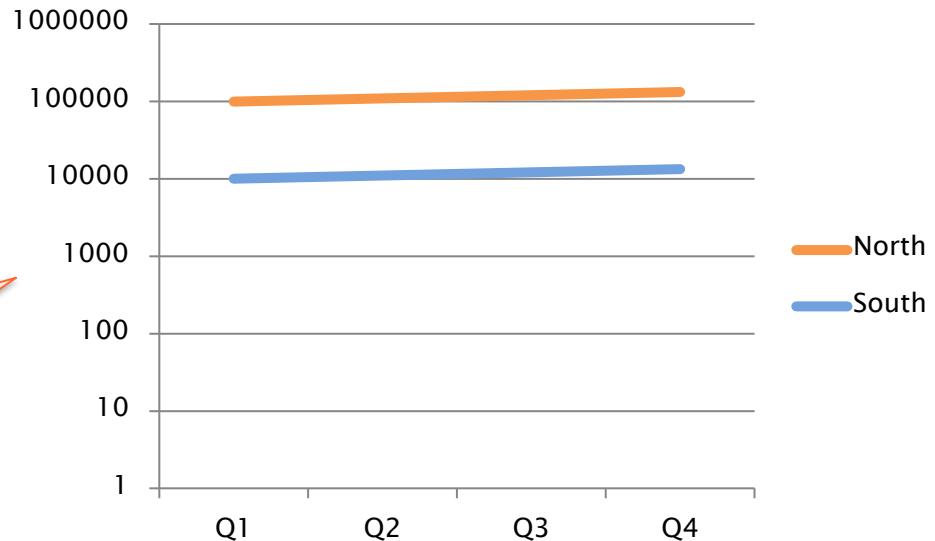


Log scale



Absolute Gains

Percentage Gains



Graph area

- Aspect ratio should not distort perception
 - ◆ Typically wider than taller
 - ◆ Scatter plots may be squared
- Grid lines must be thin and light
 - ◆ Useful to look-up values
 - ◆ Enhance comparison of values
 - ◆ Enhance perception of localized patterns

Labels

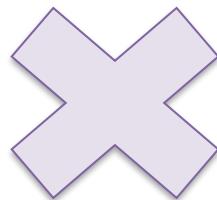
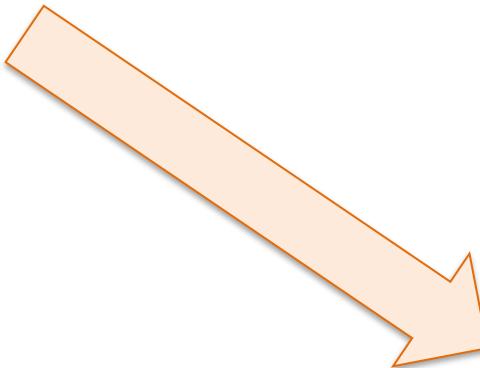
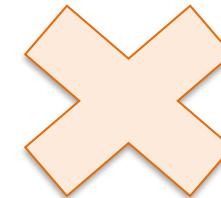
- Important elements (e.g. titles) should be prominent
 - ◆ Top
 - ◆ Larger

Gutenberg Diagram

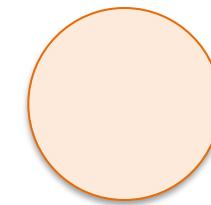
Primary area



Strong fallow area

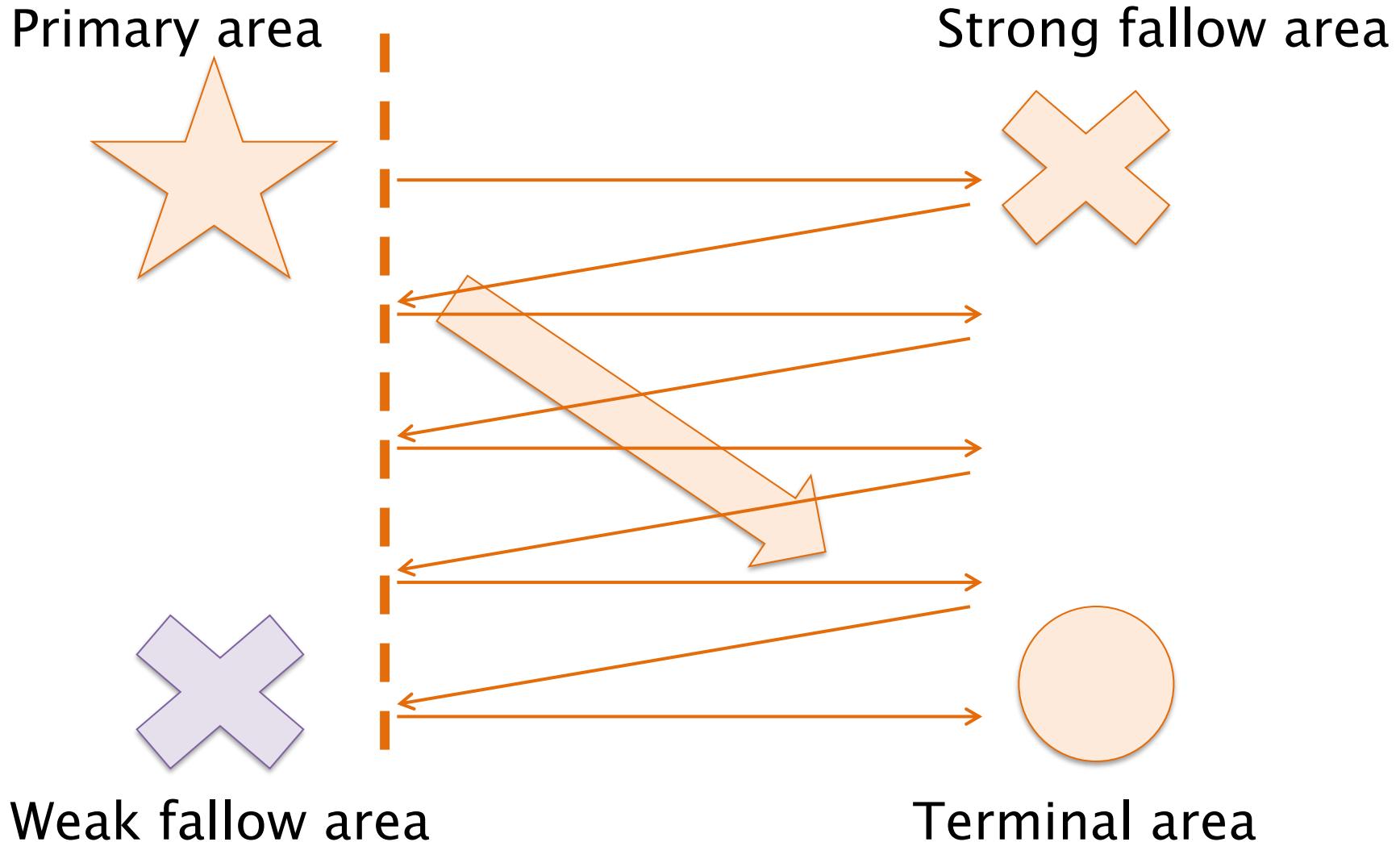


Weak fallow area



Terminal area

Gutenberg Diagram



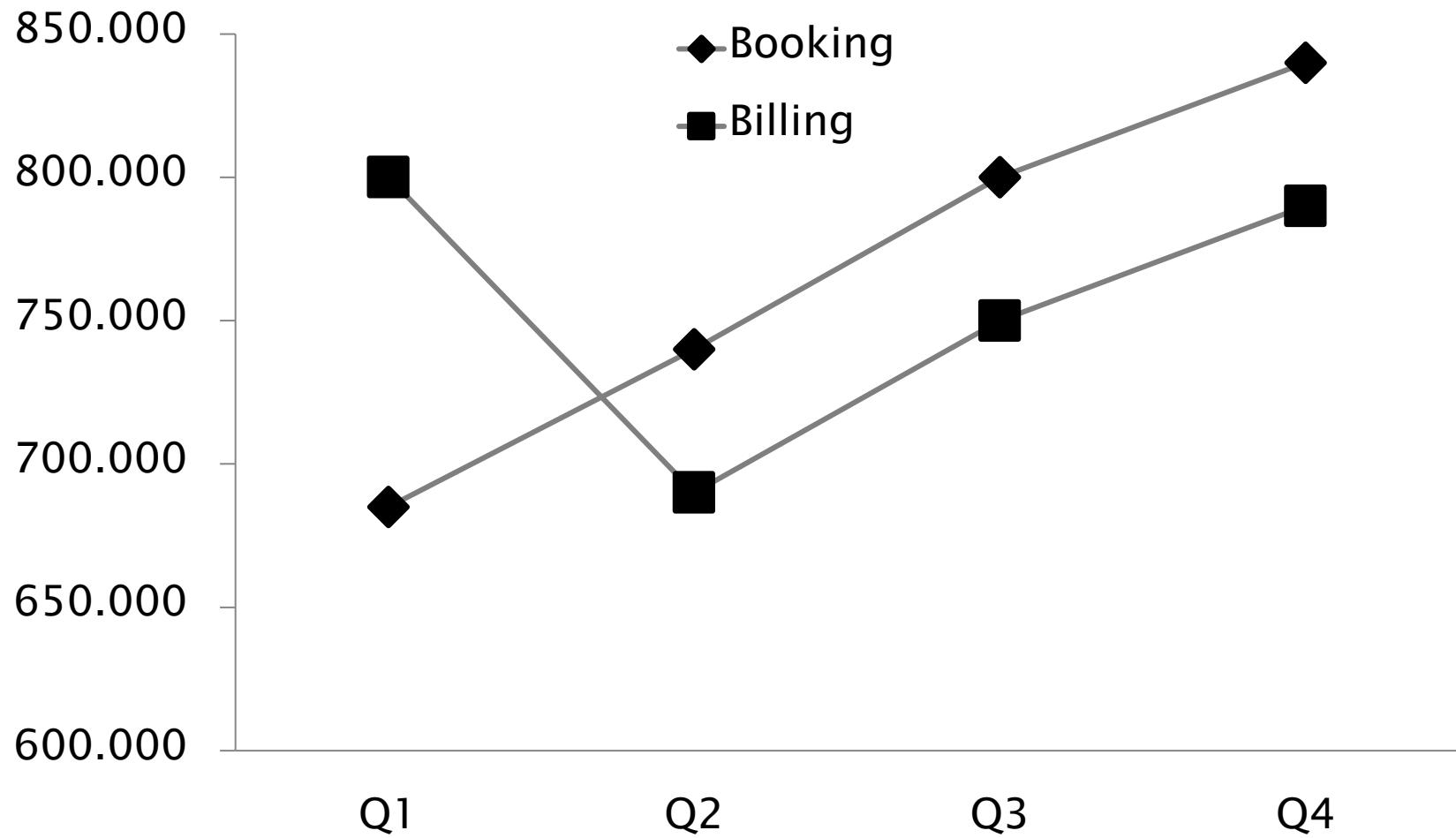
Legends

- Used for categorical attributes not associated to any axis
- As close as possible to the objects
- Less prominent than data objects
- Borders are used only when necessary to separate from other elements

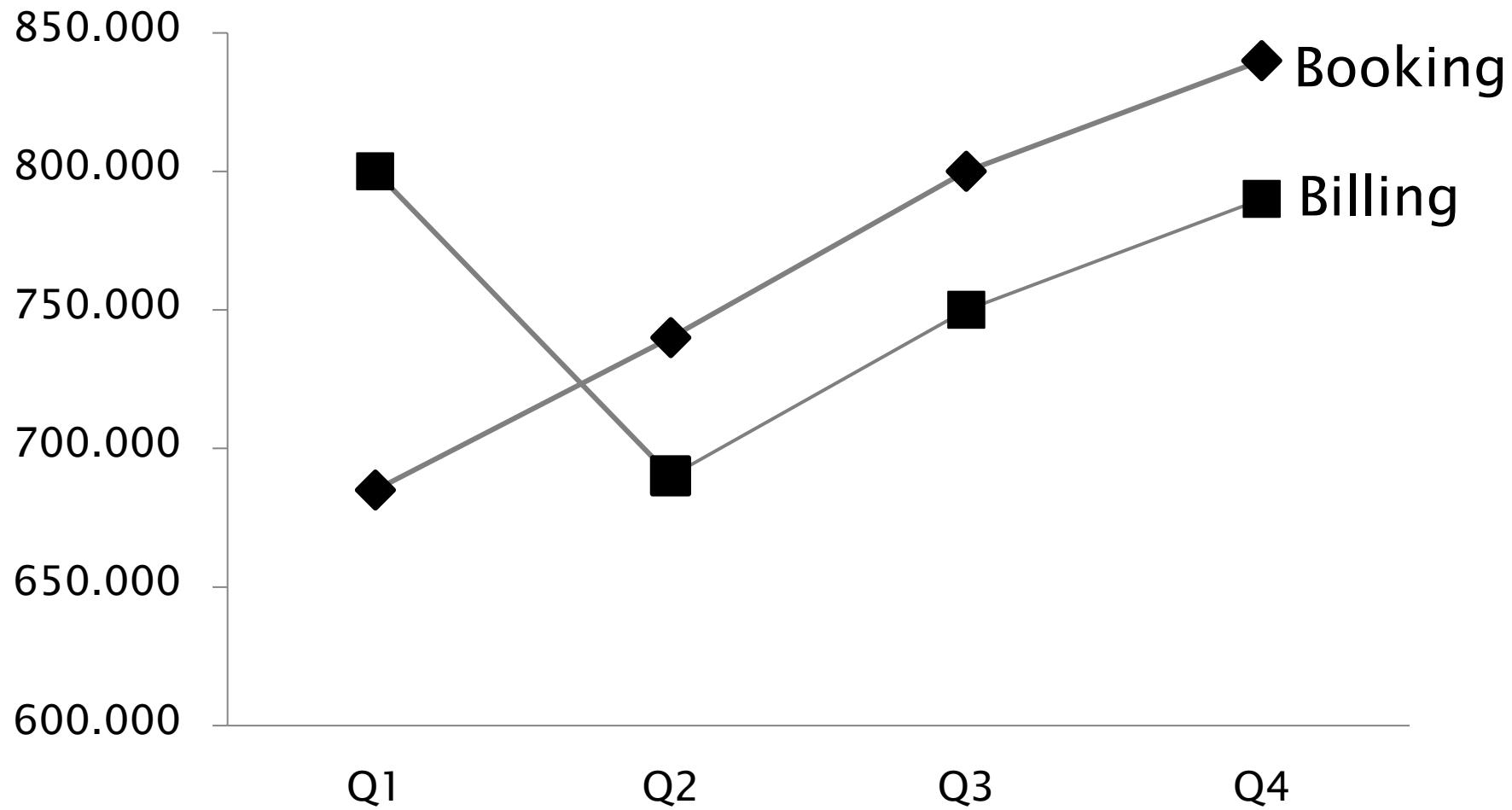
Legends

- Text should be as close as possible to the object it complements
 - ◆ Prefer direct labeling to separate legends
- Number of categorical subdivisions
 - ◆ Perceptual limit is between 5 and 8
 - ◆ Limit is independent of the visual attribute used to encode it
 - ◆ Joint use of attributes ease discrimination

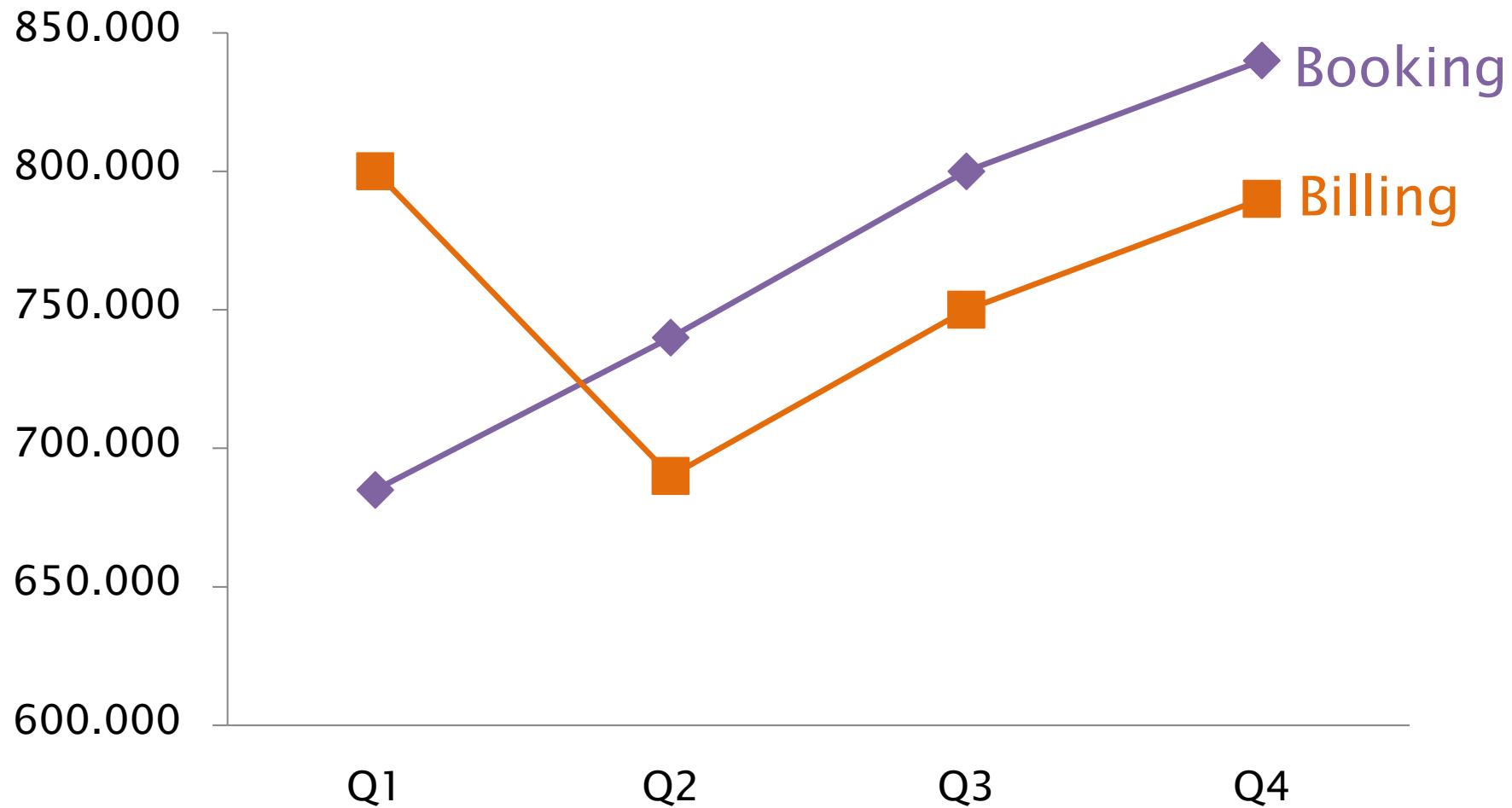
Legend



Direct labeling

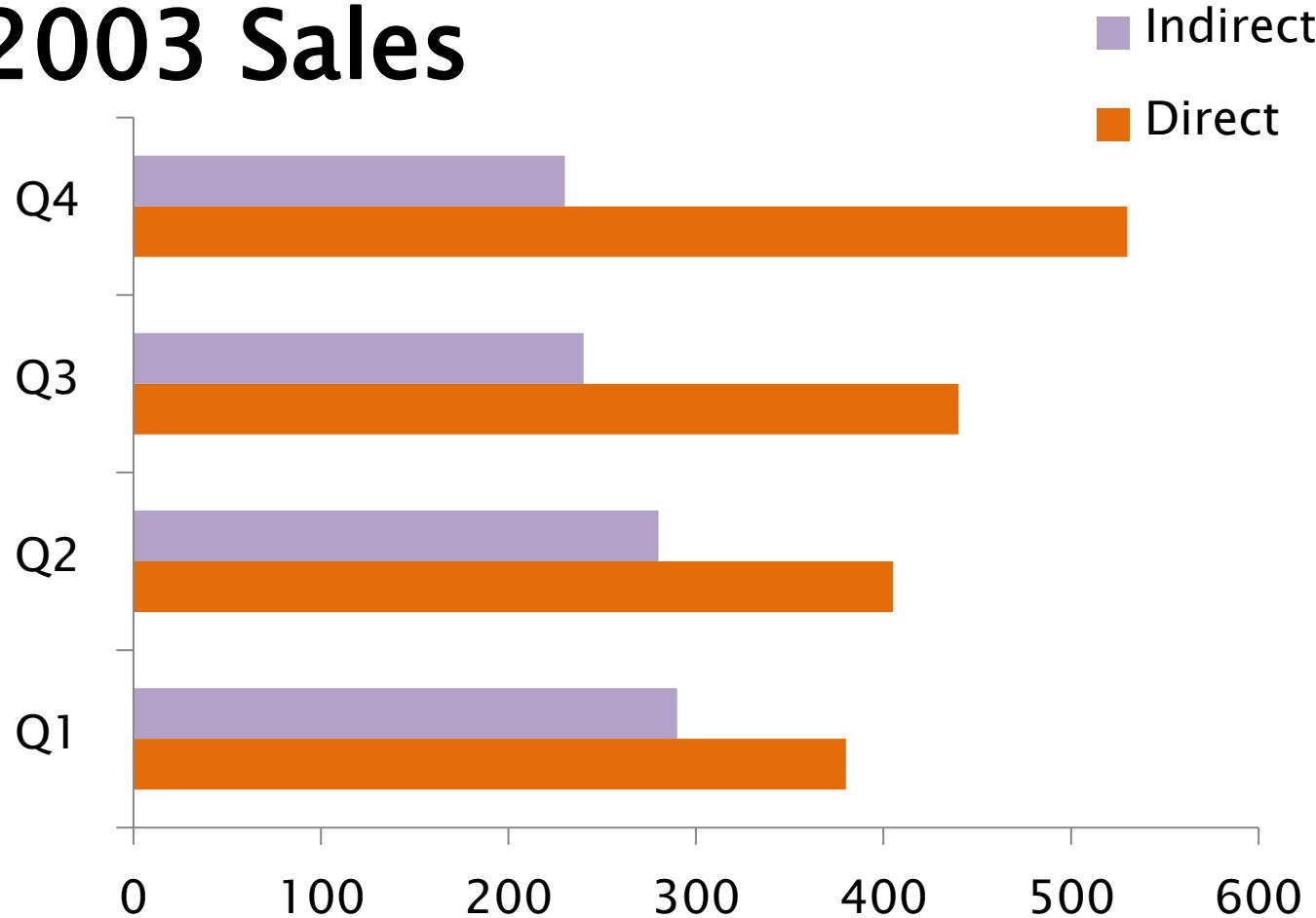


Direct labeling and color



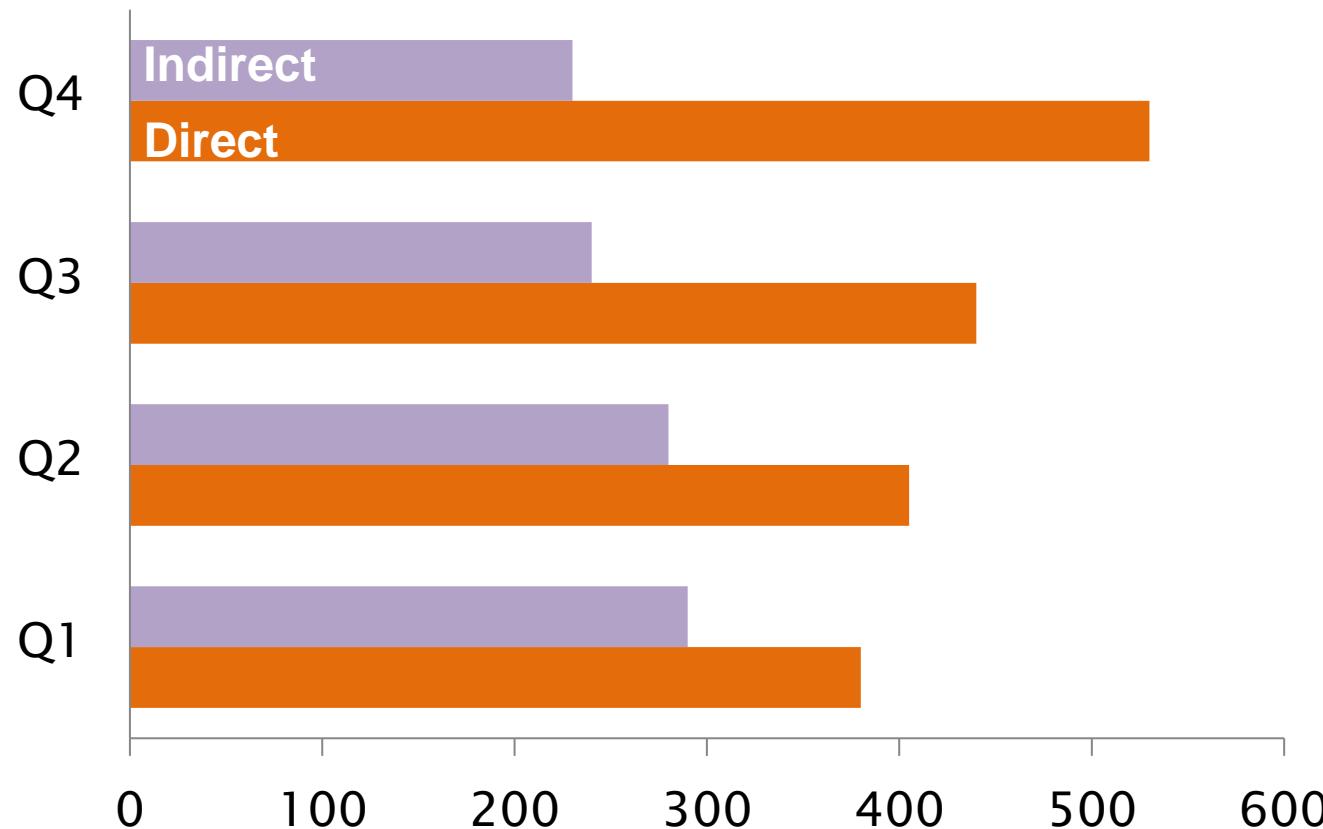
Legend

2003 Sales



Direct labeling

2003 Sales



Reference lines and regions

- Reference lines support an easy comparison to a given value
 - ◆ Mean
 - ◆ Threshold
- Reference regions allow comparison with several values
 - ◆ Use background color

VISUAL RELATIONSHIPS

Data Visualization

Understanding

Information Visualization

Visual Patterns, Trends, Exceptions

Quantitative Reasoning

Quantitative Relationship & Comparison

Visual Perception

Visual Properties & Objects

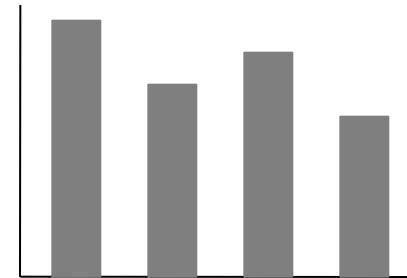
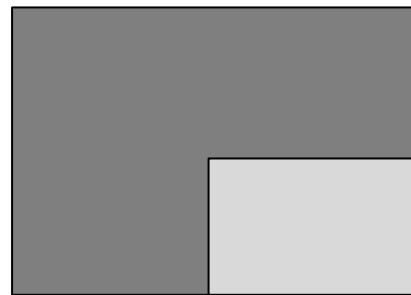
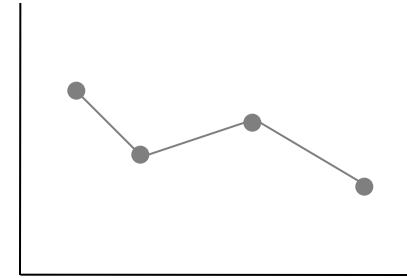
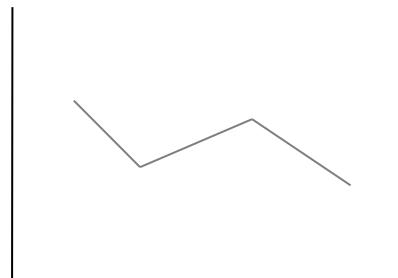
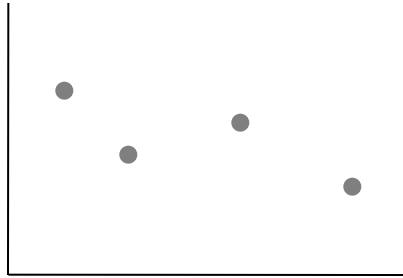
Data

Representation/Encoding

Relationships

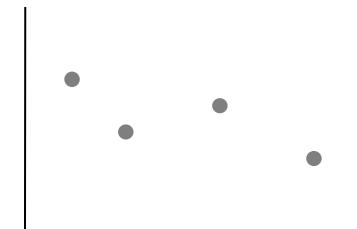
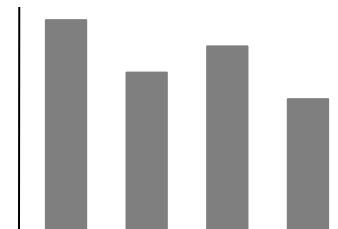
- Within a category
 - ◆ Nominal comparison
 - ◆ Ranking
 - ◆ Part-to-whole
 - ◆ Distribution
- Between measures
 - ◆ Time series
 - ◆ Deviation
 - ◆ Correlation

Quantitative encoding

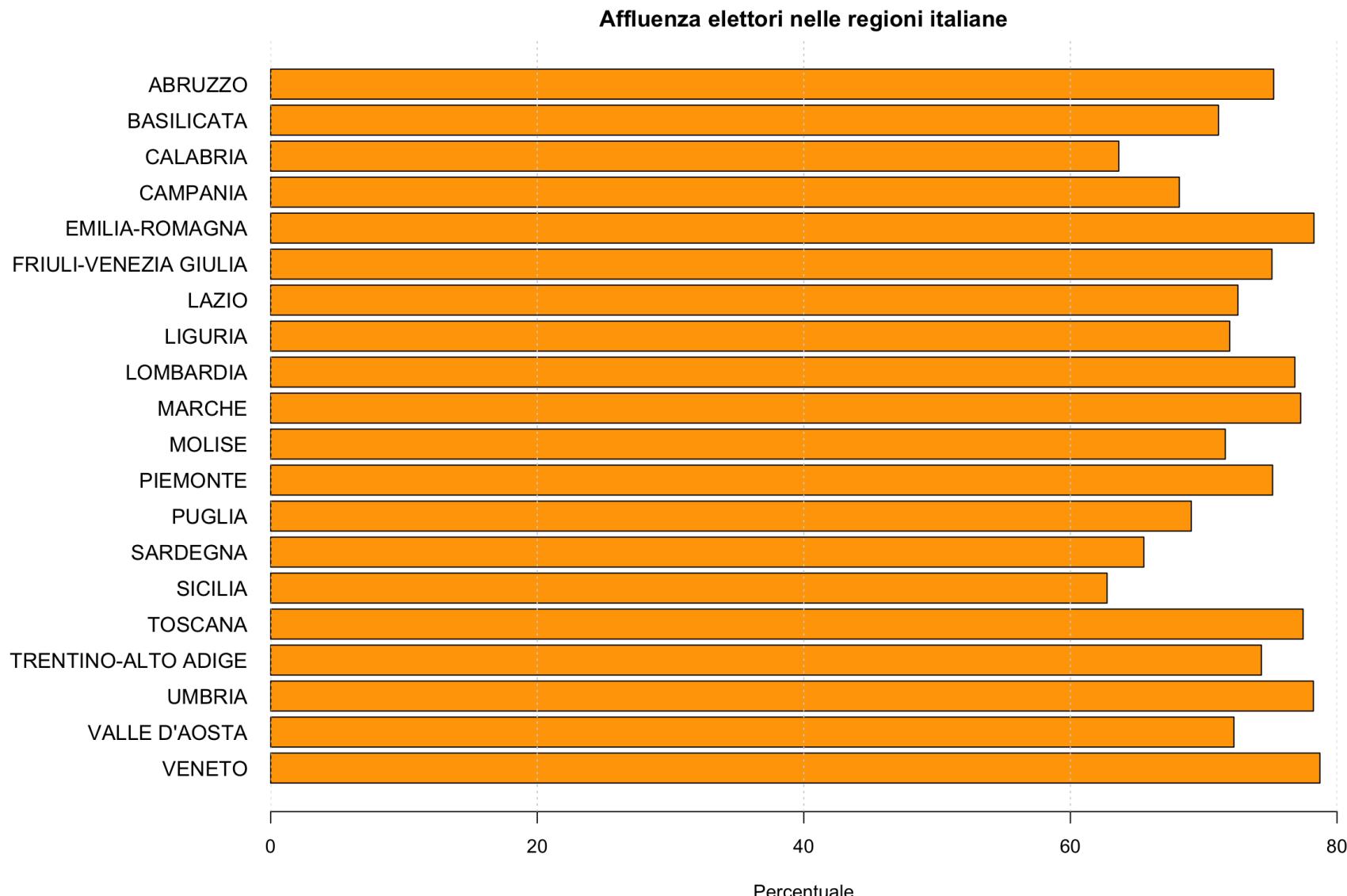


Nominal comparison

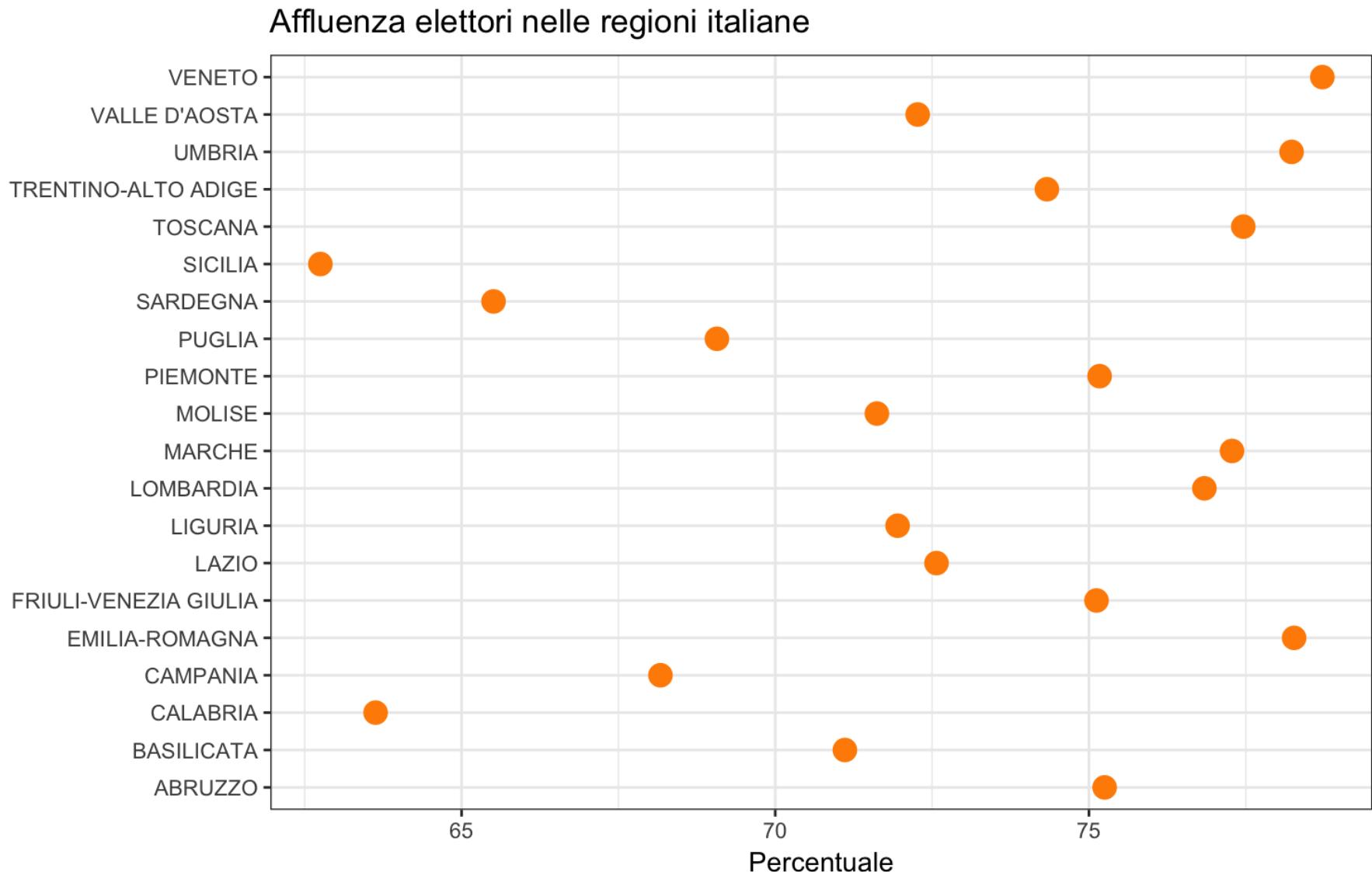
- Compare quantitative values corresponding to categorical levels
 - ◆ Small differences are difficult to see
 - Non zero-based scale can emphasize
 - ◆ Dot plots can be used for small differences
 - They do not require zero based scale



Comparison – Barplot



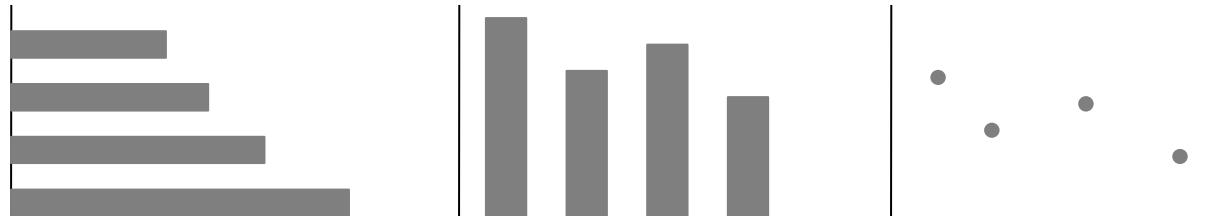
Comparison – Dot plot



Ranking

- Same type as nominal comparison
- Pay attention to order

Purpose	Sort order	Bars orientation
Highlight the highest value	Descending	H: highest on top V: highest on left
Highlight the lowest value	Ascending	H: lowest on top V: lowest on left

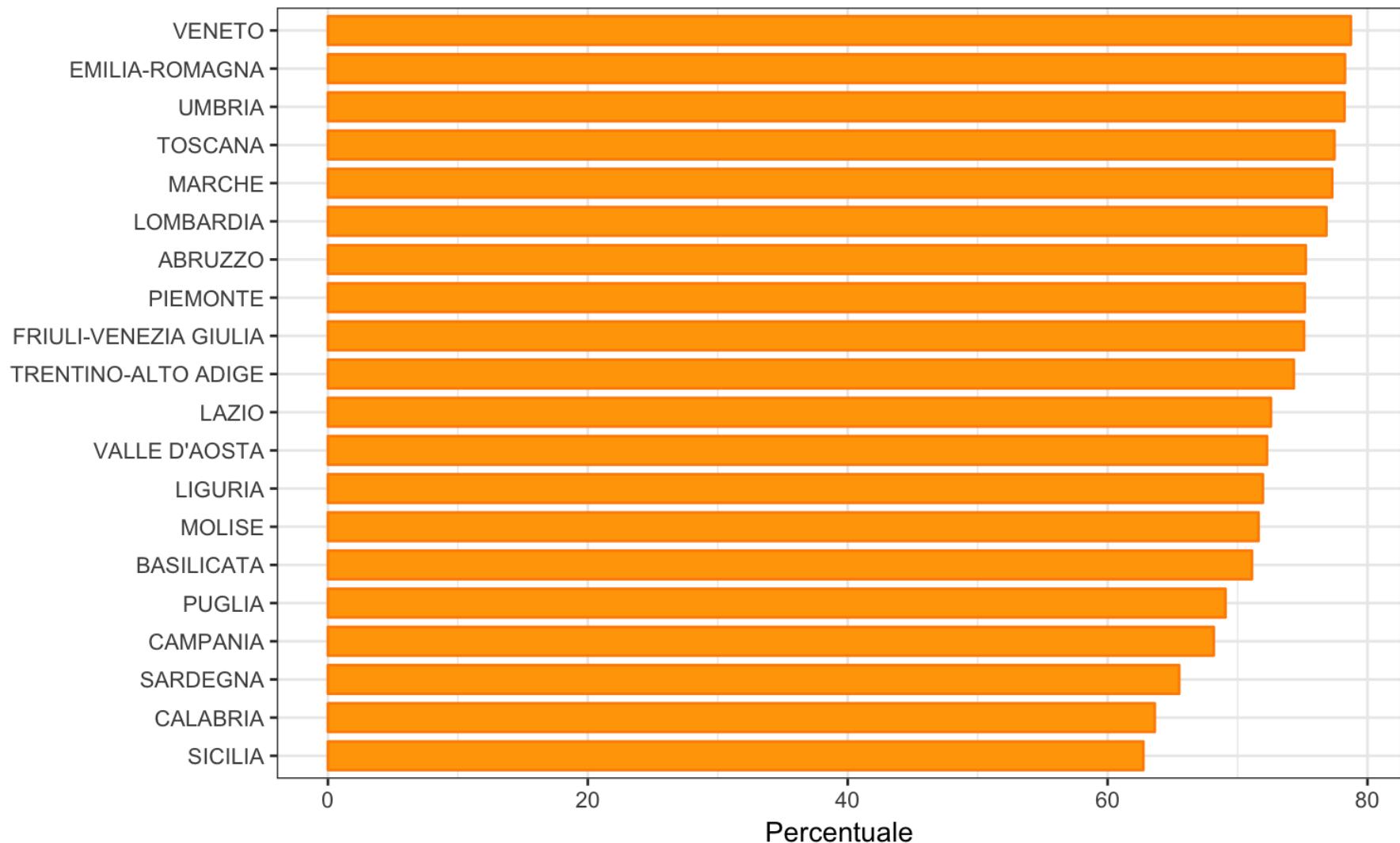


Ranking

- Bar graphs
- Dot plot
 - ◆ Allow non zero-based axes
- Line charts
 - ◆ Show evolution in time

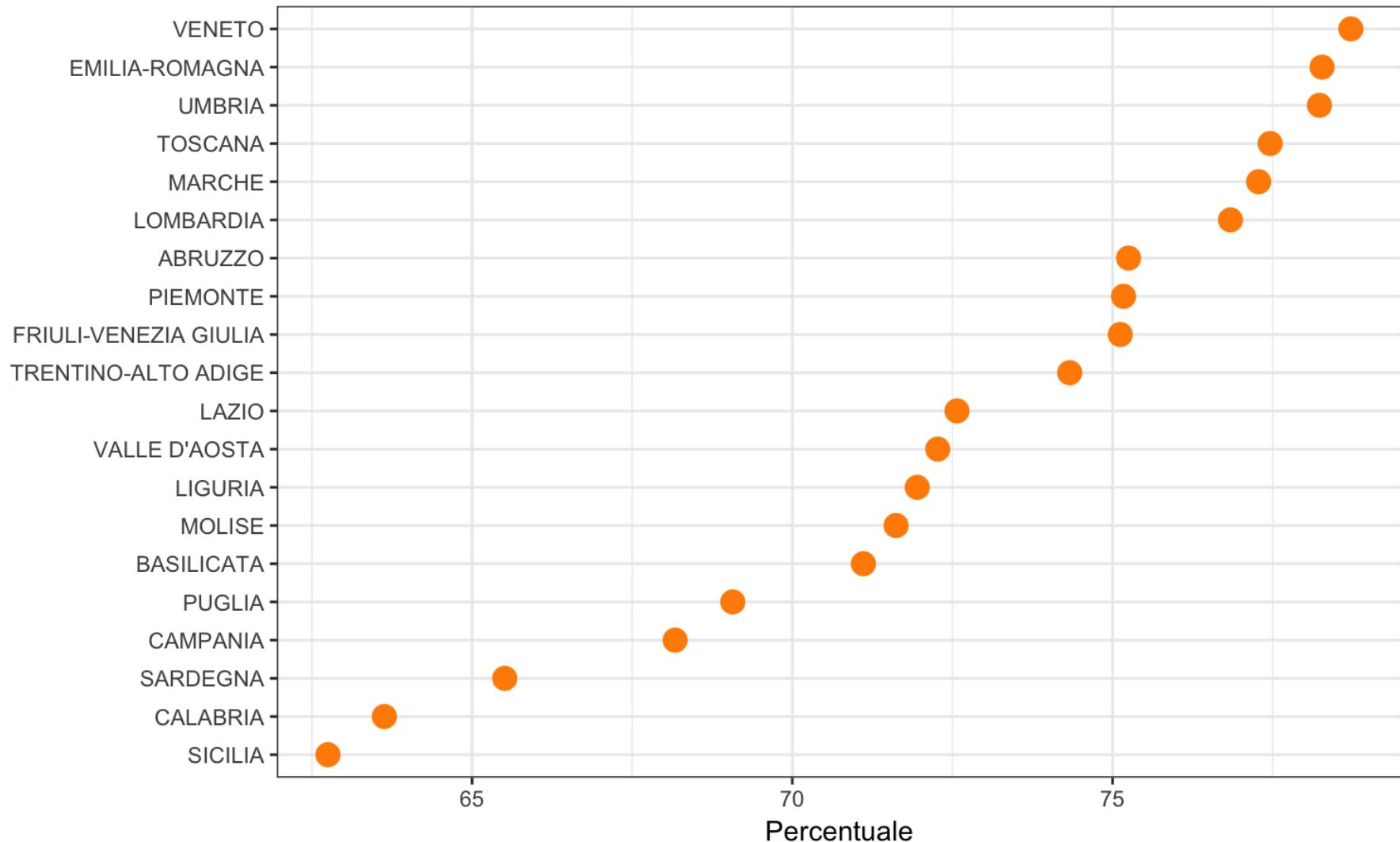
Ranking – Barplot

Affluenza elettori nelle regioni italiane



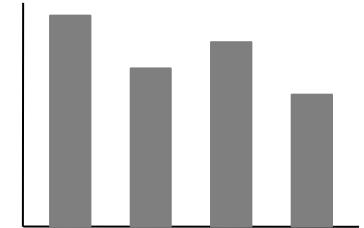
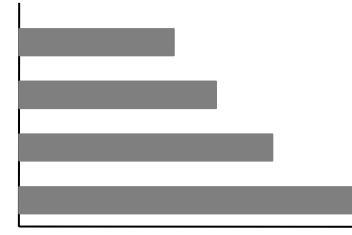
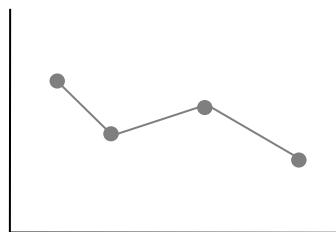
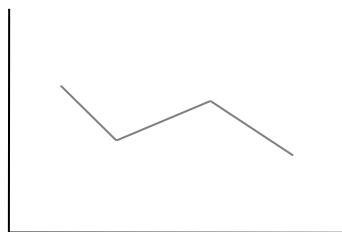
Ranking – Dot plot

Affluenza elettori nelle regioni italiane



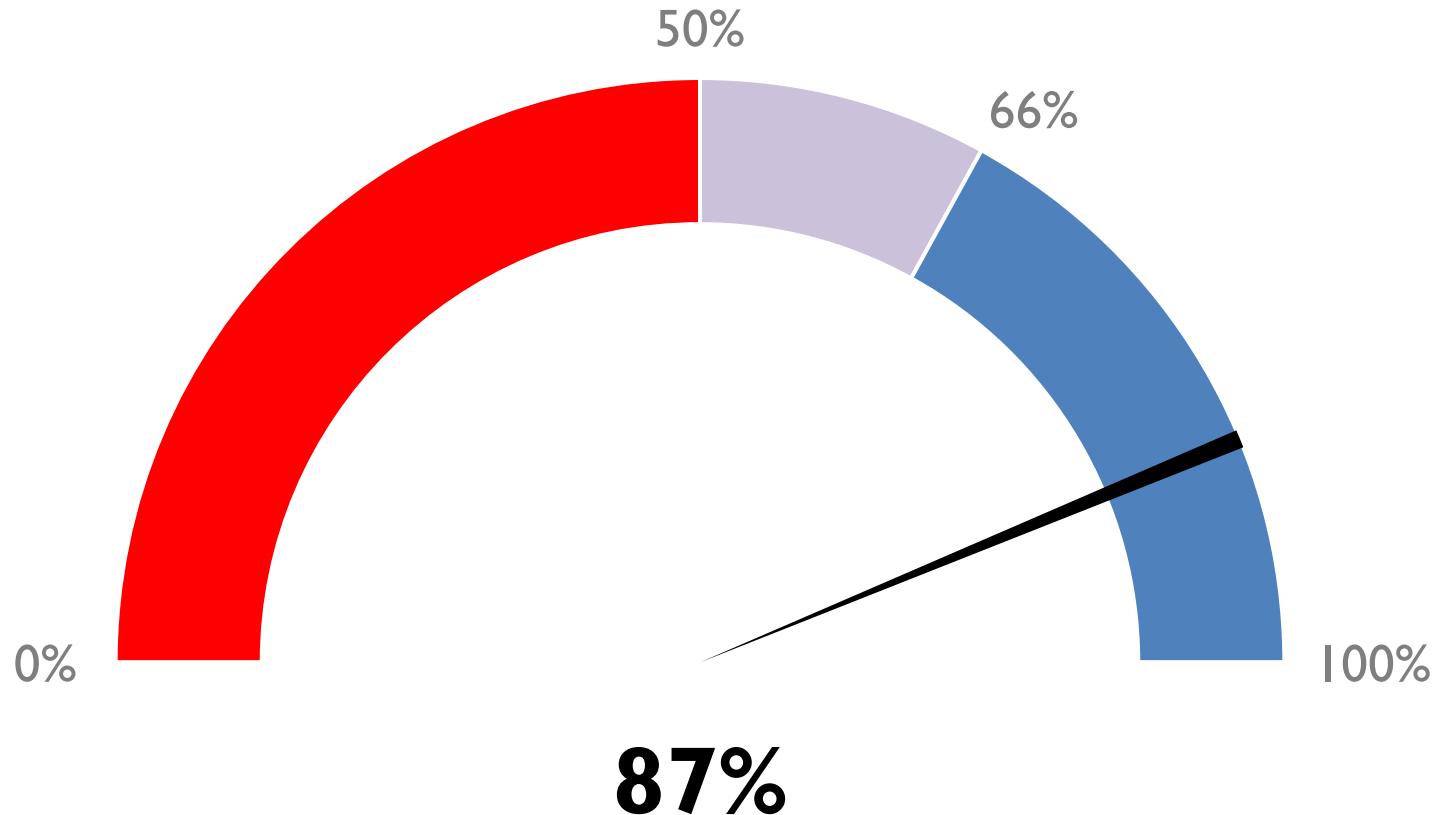
Deviation

- To what degree one or more sets of values differ in relation to primary values.
 - ◆ Points
 - ◆ Gauge
 - ◆ Bars
 - ◆ Bullet

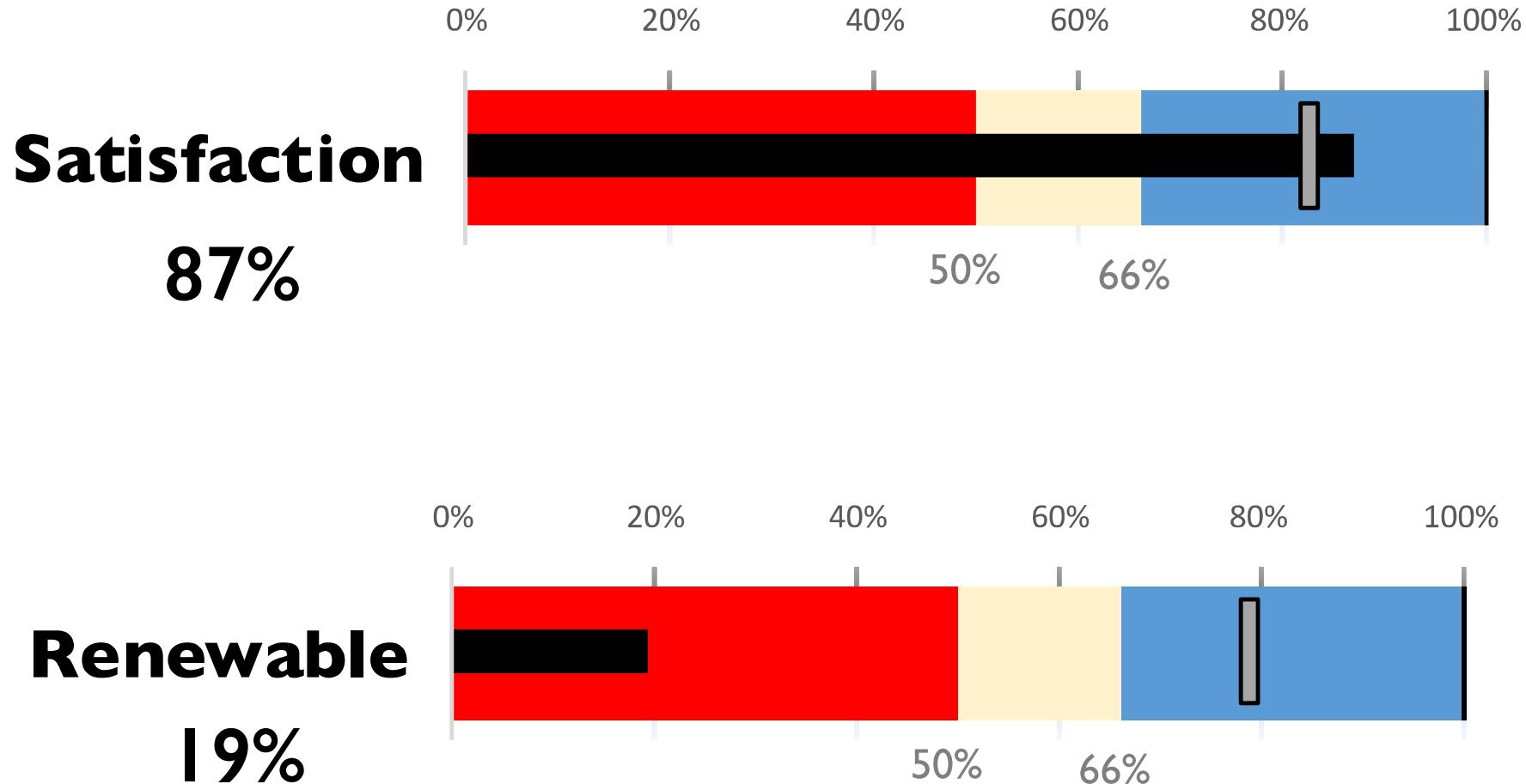


Angle + Position – Gauge

Satisfaction



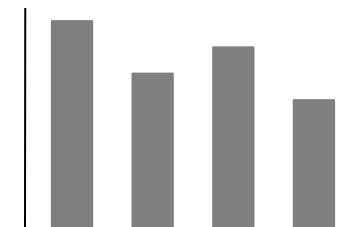
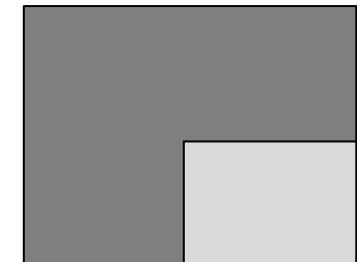
Length+Position- Bullet Graph



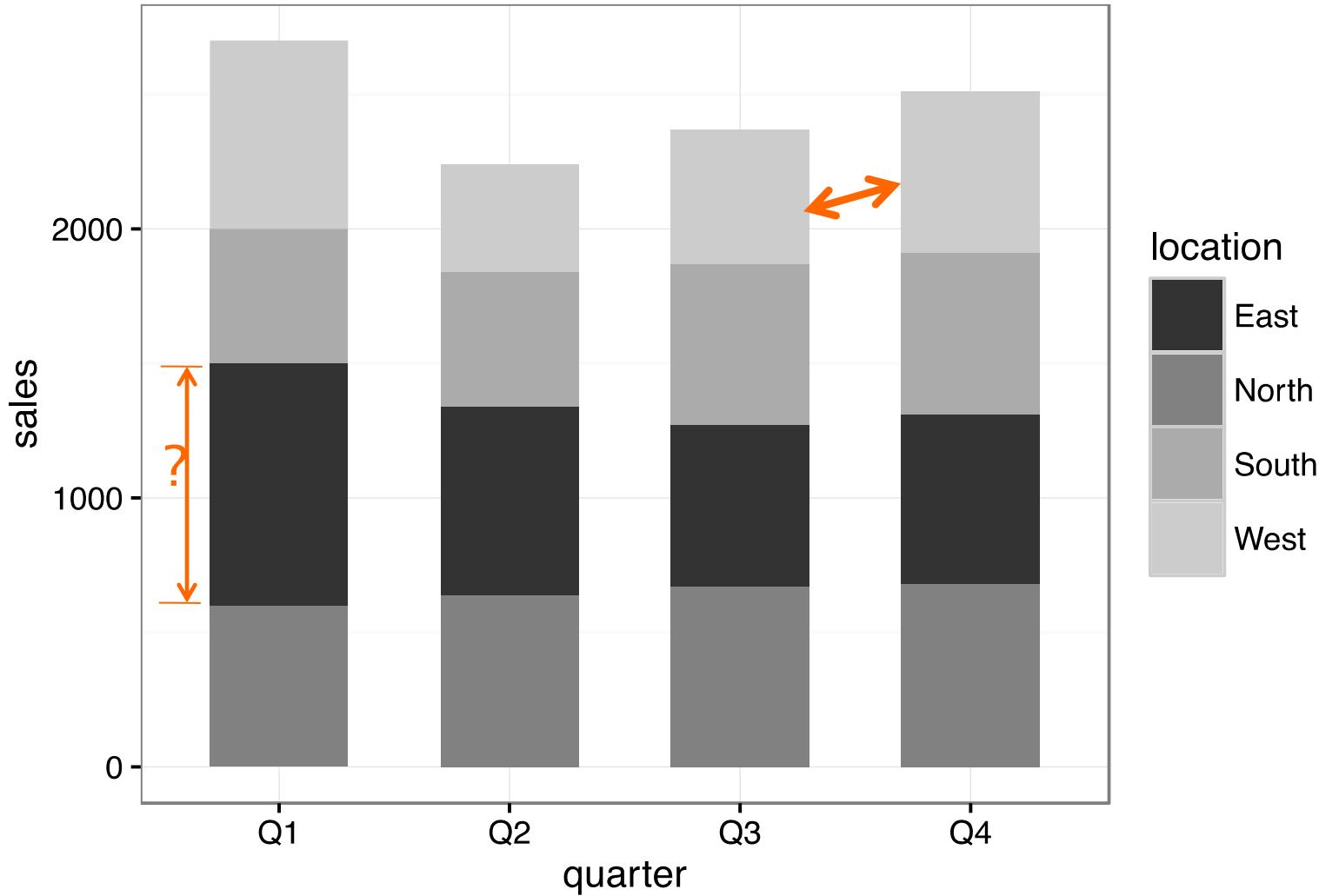
https://www.perceptualedge.com/articles/misc/Bullet_Graph_Design_Spec.pdf

Proportion (Part-to-whole)

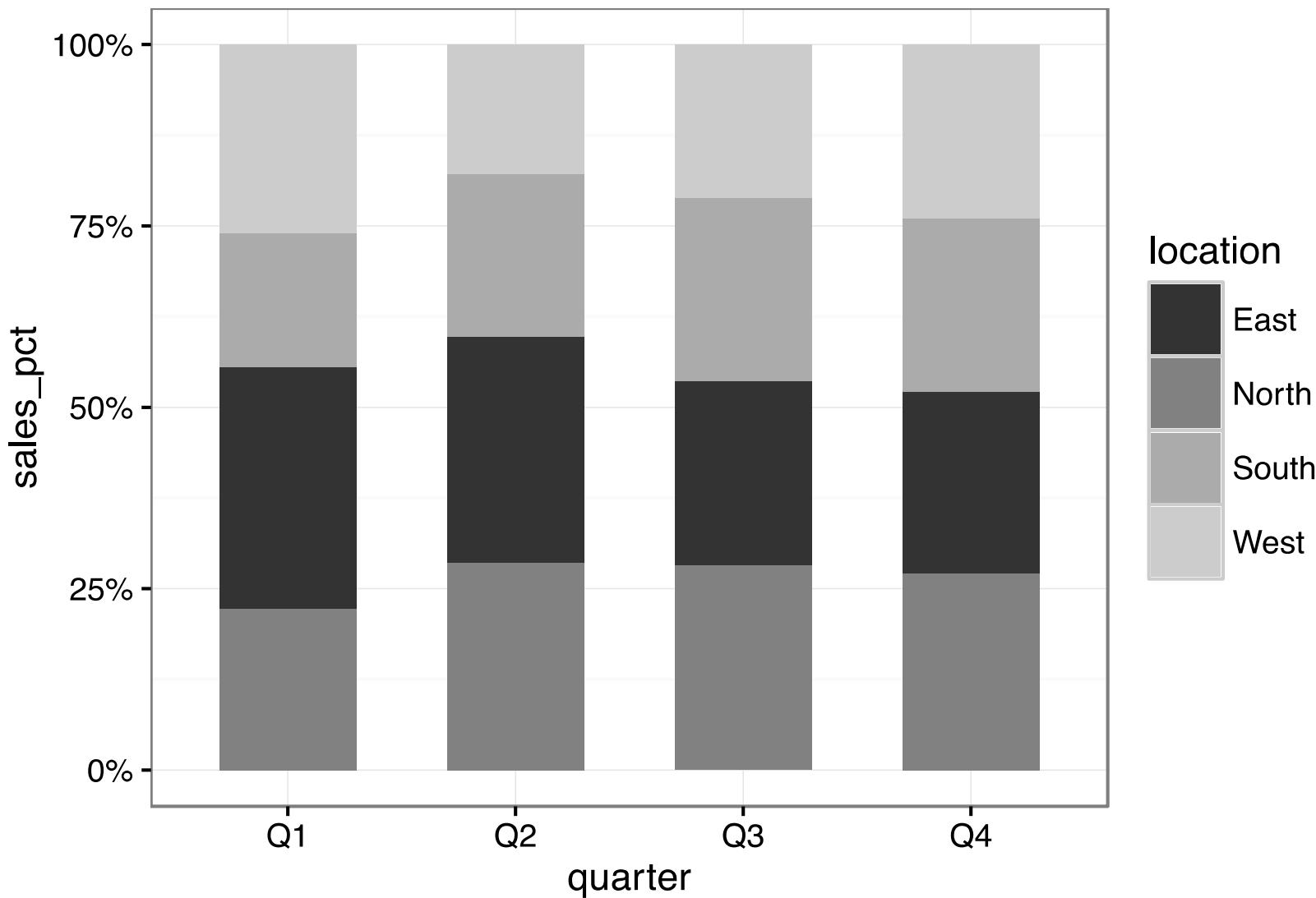
- Best unit: percentage
- Stacked bar graph
 - ◆ Difficult to read individual values
- Stacked area
- Treemap
- Gridplot
- Pie / Donut
- Marimekko



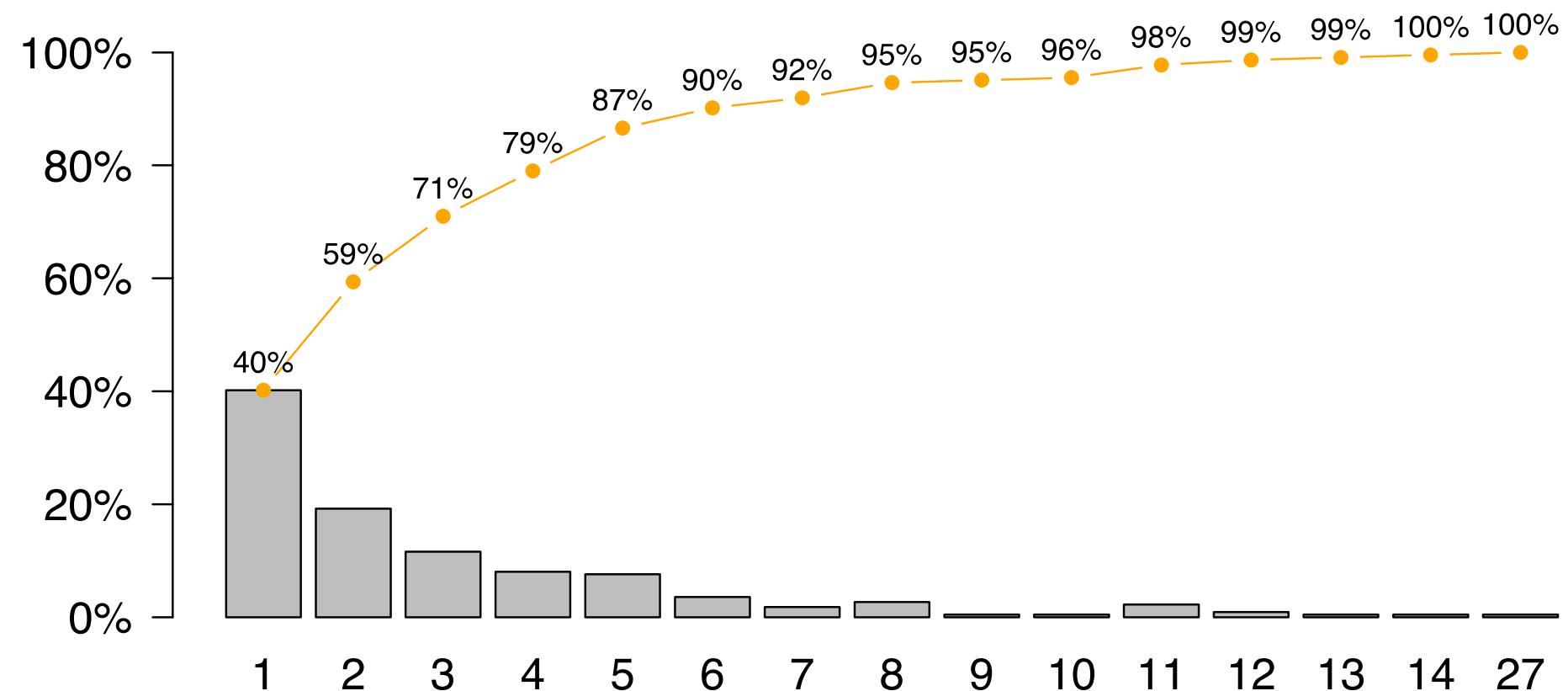
Stacked bar graph



Stacked bar graph percentage



Pareto chart

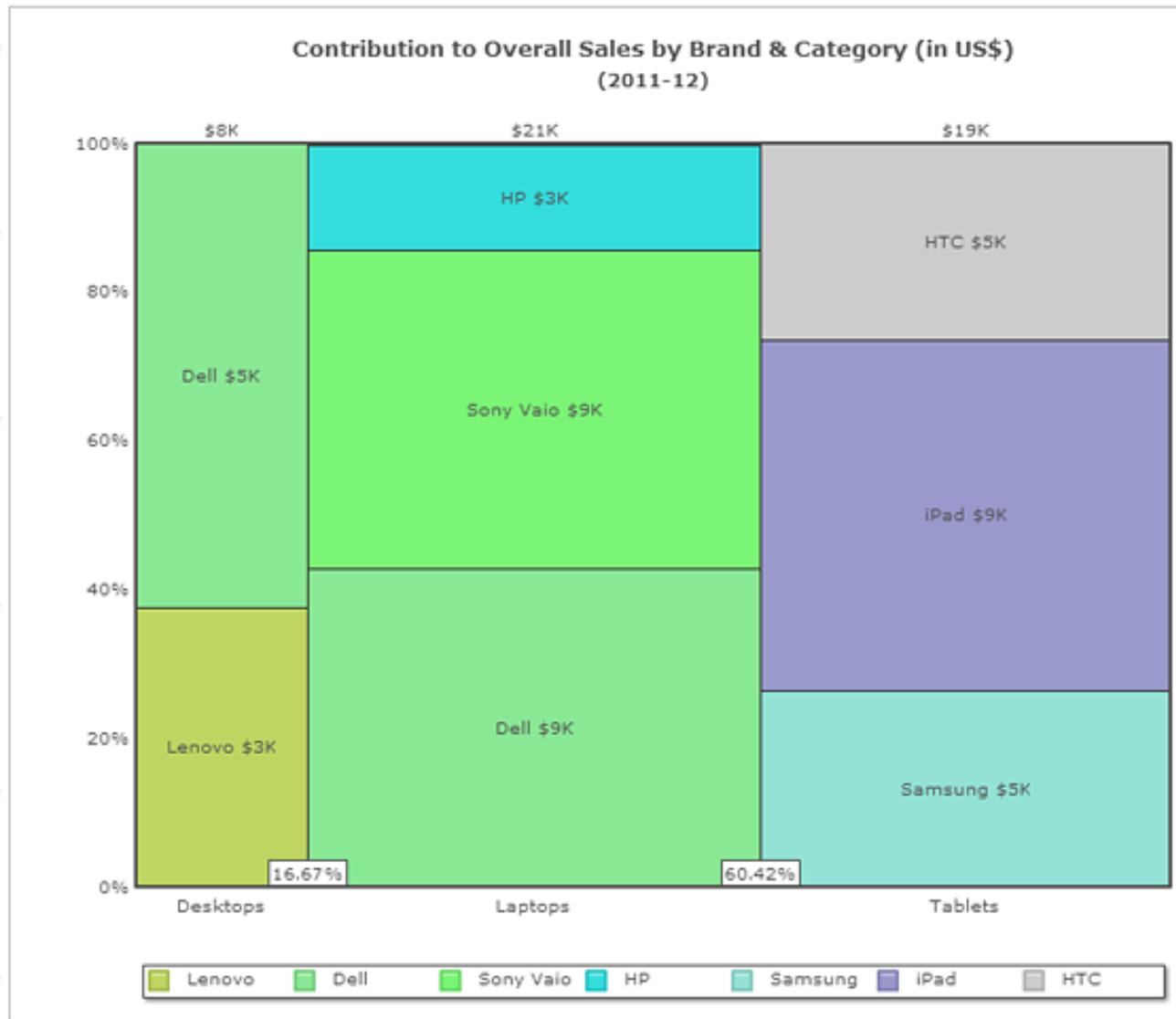


Treemap

proporzione



Marimekko Chart



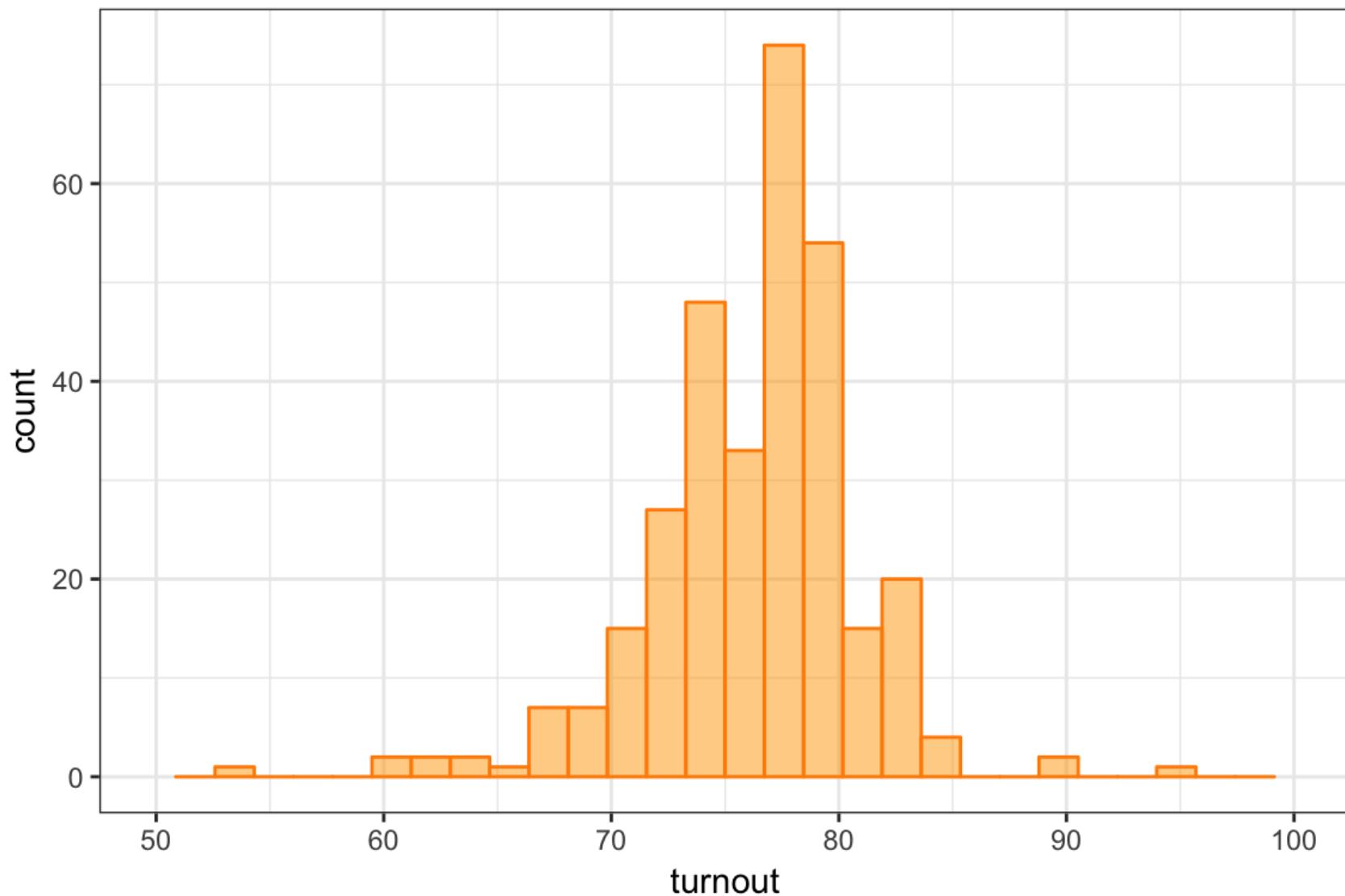
Distribution

- Two main types
 - ◆ Show distribution of single set of values
 - ◆ Show and compare two or more distributions

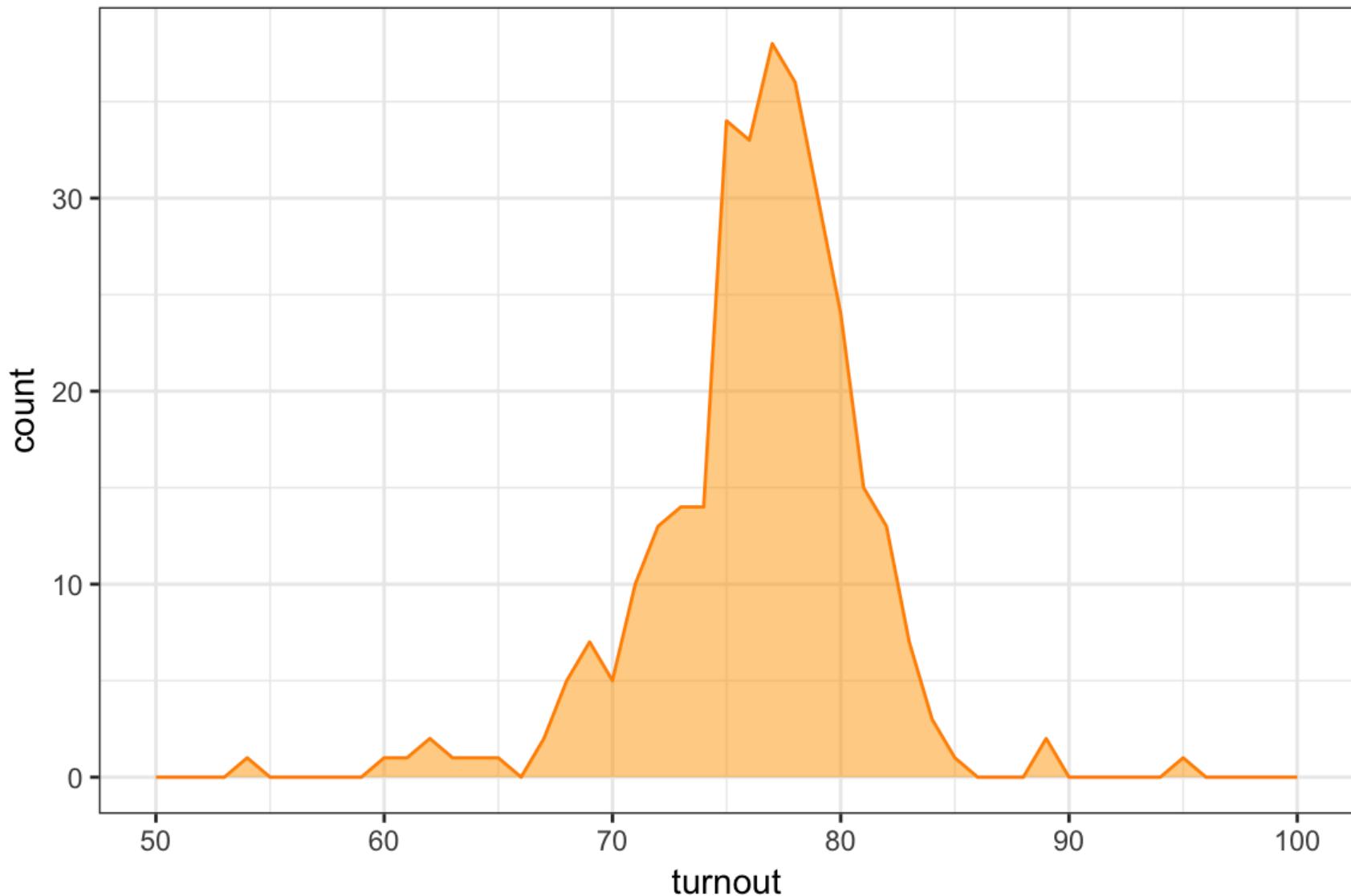
Single distribution

- Histogram
 - ◆ Vertical bar graph
 - ◆ Frequency for subdivision
 - Quantitative ranges
 - Categories
 - ◆ Emphasis on number of occurrences
- Frequency polygon
 - ◆ Line graphs
 - ◆ Frequency density function
 - ◆ Emphasis on the shape of the distribution
- Boxplot
 - ◆ Summary

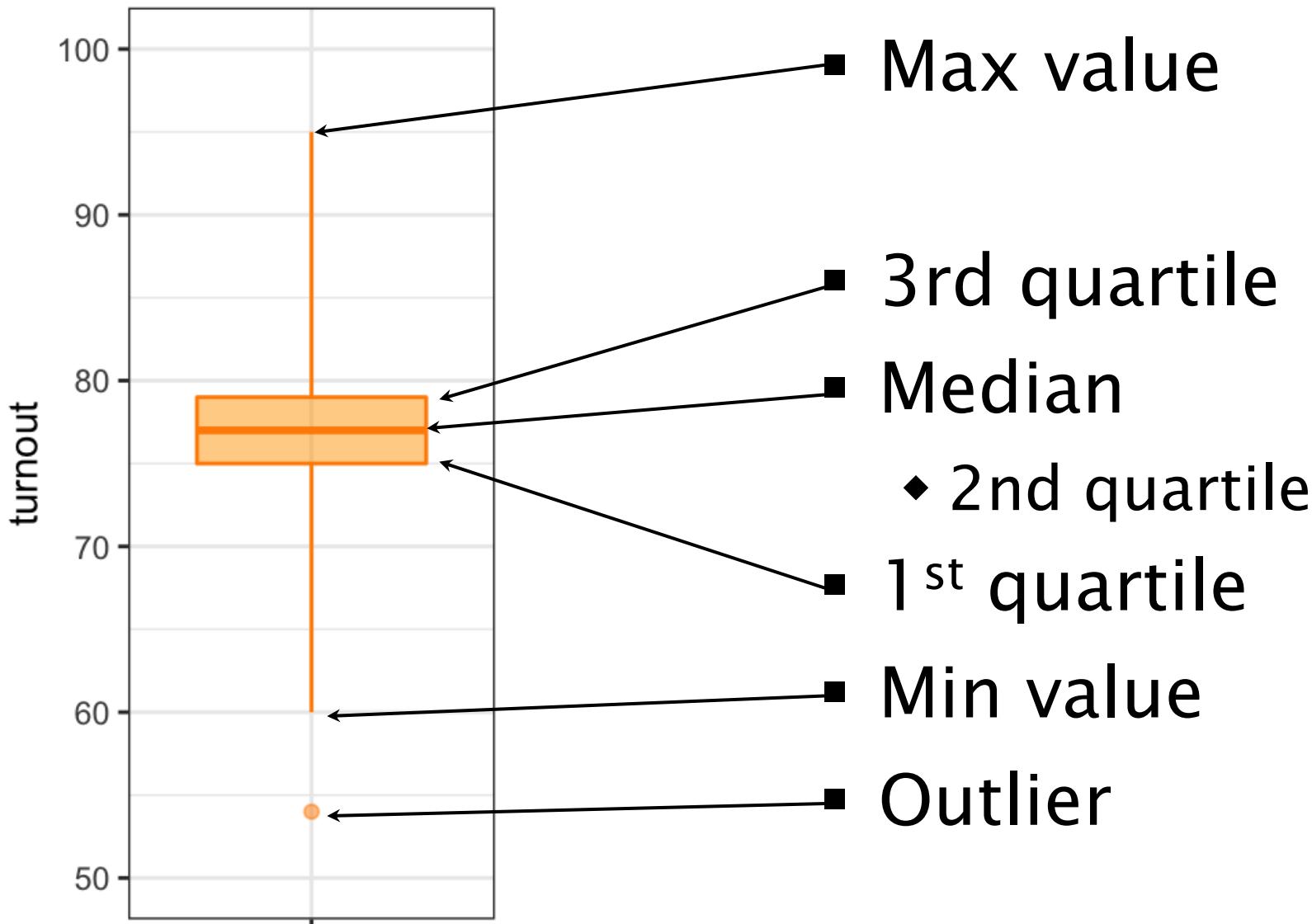
Histogram



Frequency polygon



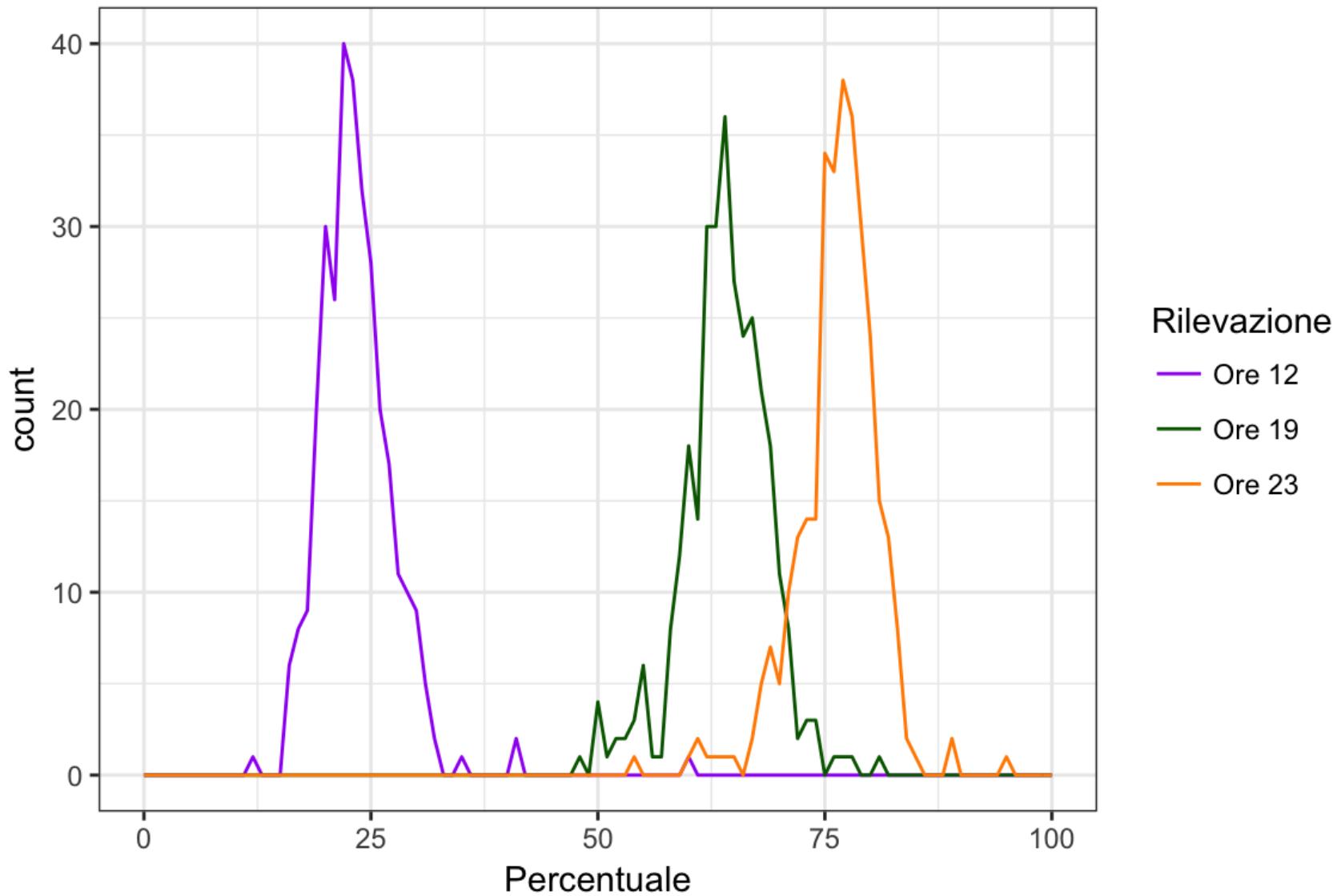
Boxplot



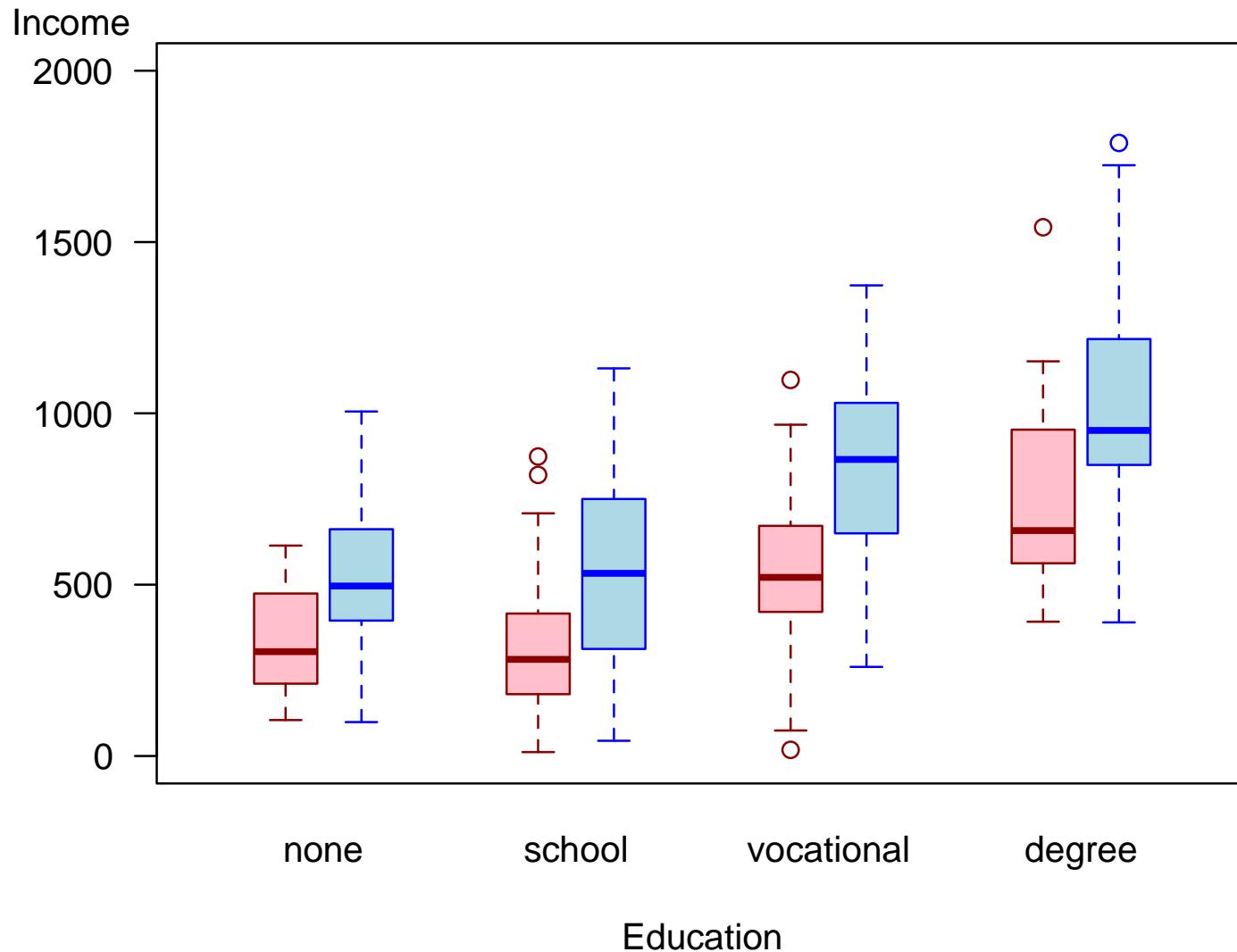
Multiple distribution

- Histogram is not suitable
- Frequency polygon
 - ◆ Line graphs
 - ◆ Frequency density function
- Boxplot
 - ◆ Summary
 - ◆ Less distracting with high number of categories

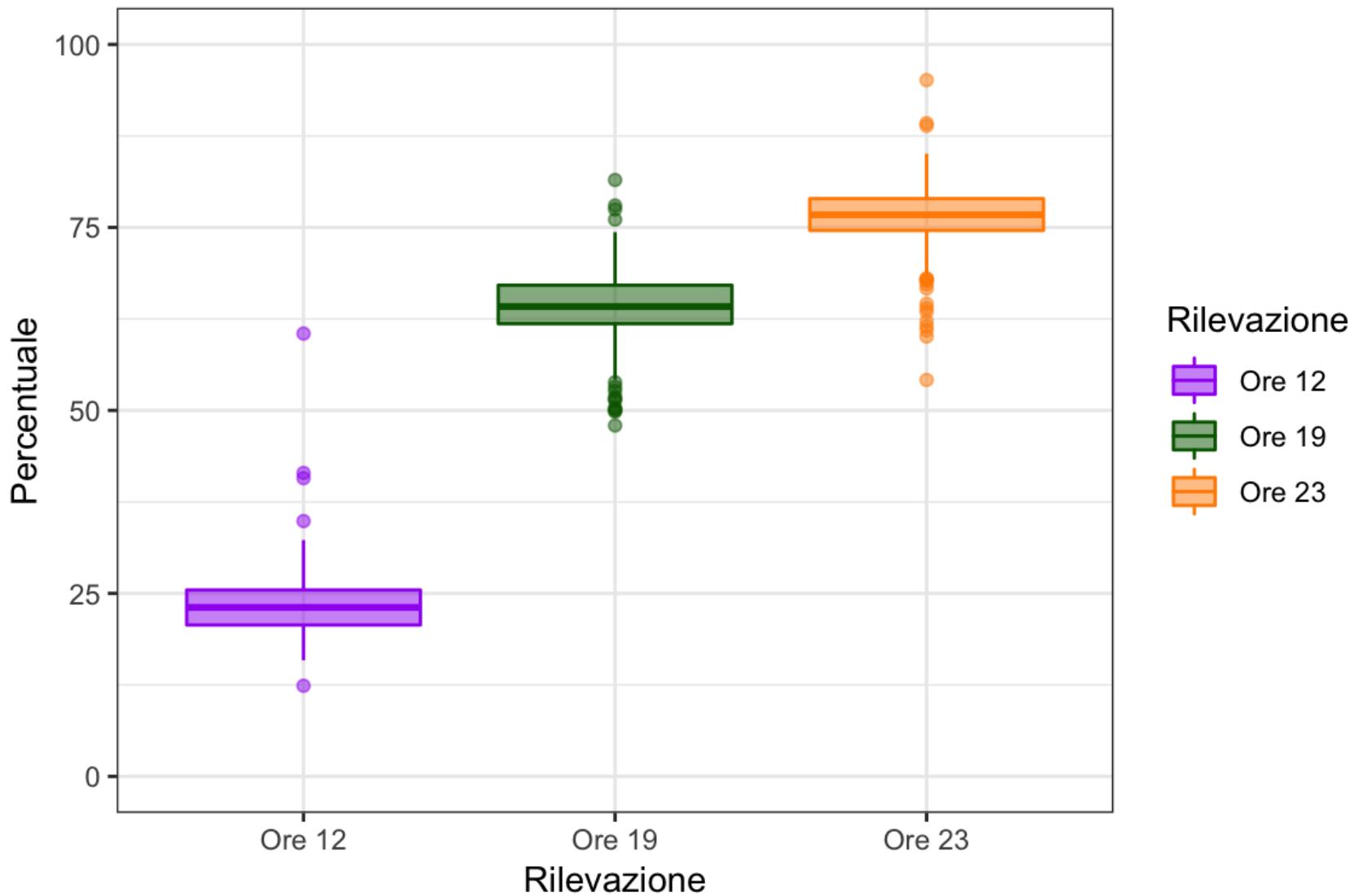
Multiple polygons



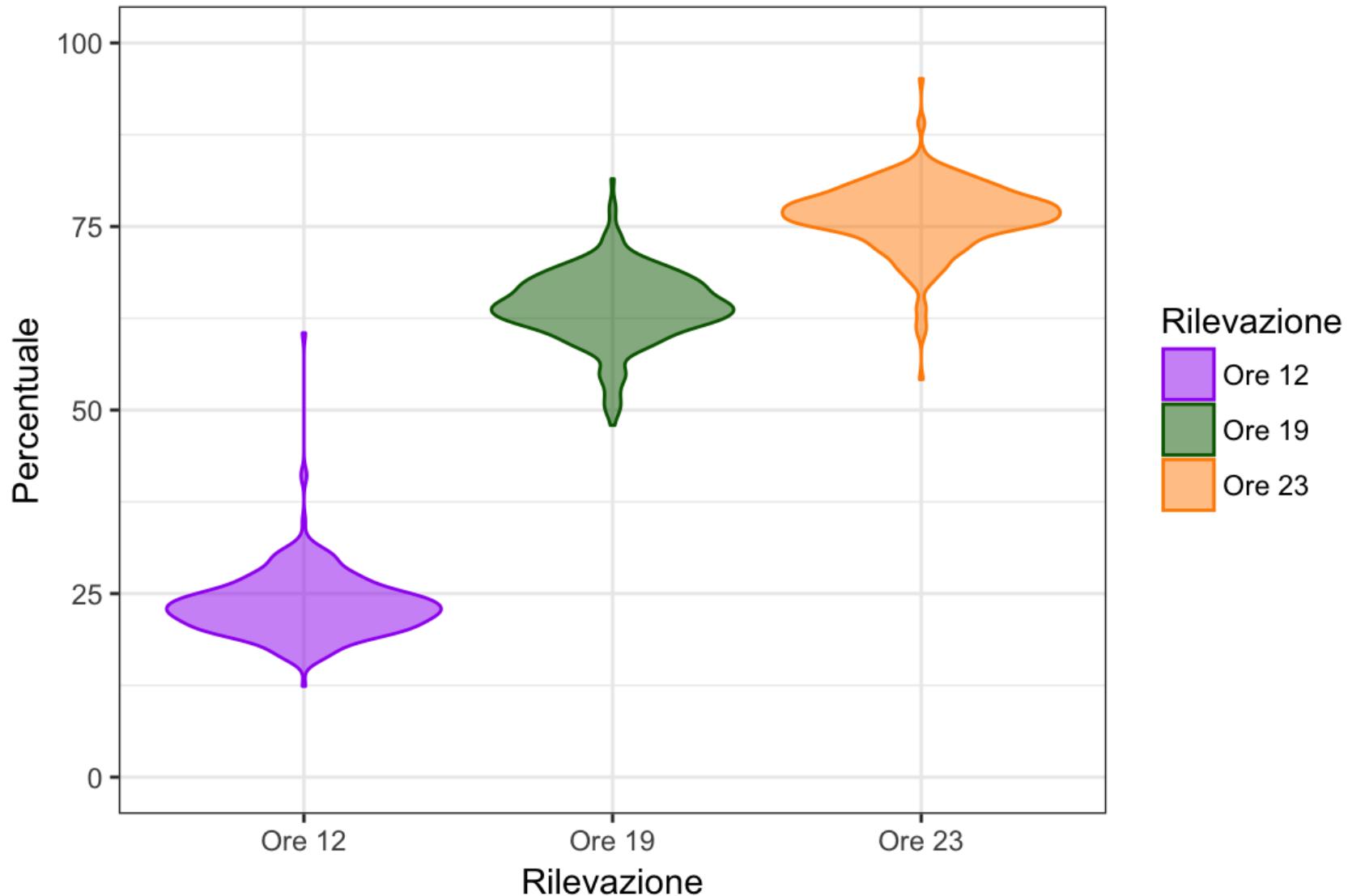
Box plot



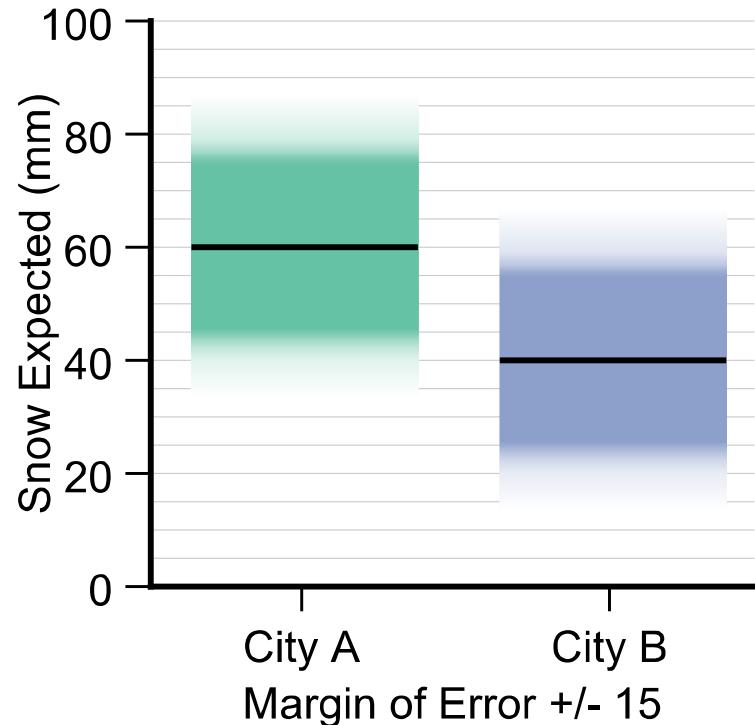
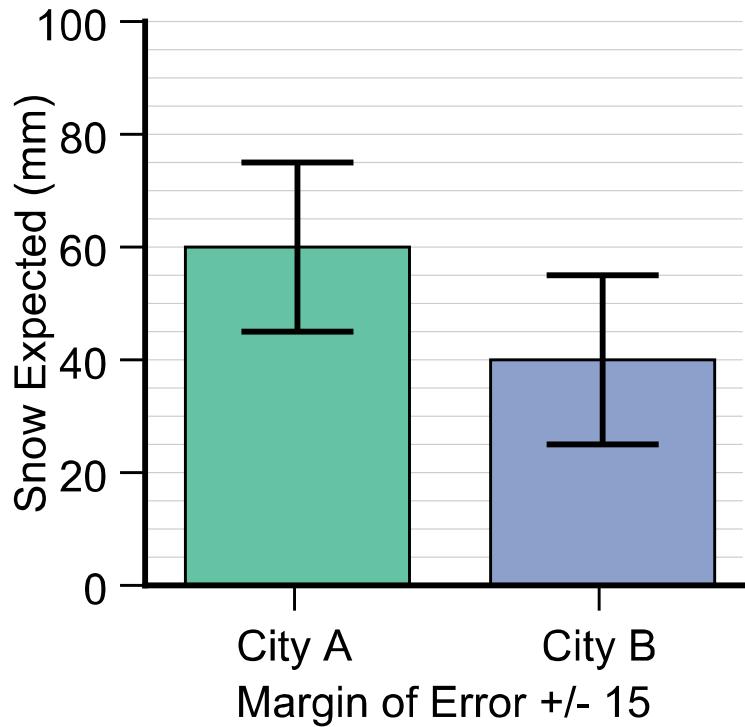
Box plot



Violin plot

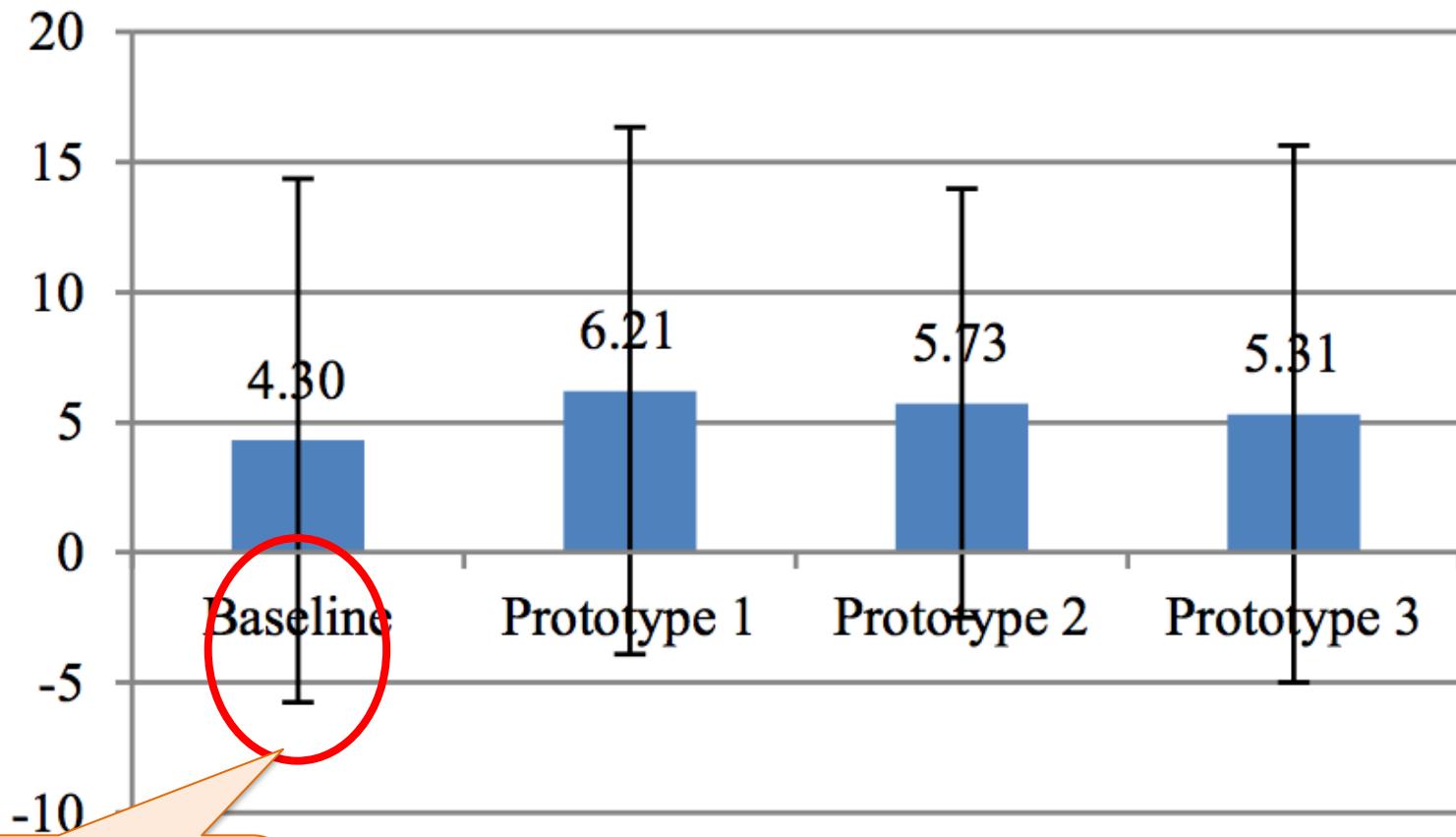


Confidence Intervals



Error Bars Considered Harmful: Exploring Alternate Encodings for Mean and Error
Michael Correll, and Michael Gleicher
IEEE Transactions on Visualization and Computer Graphics, Dec. 2014

Interval may be Asymmetric



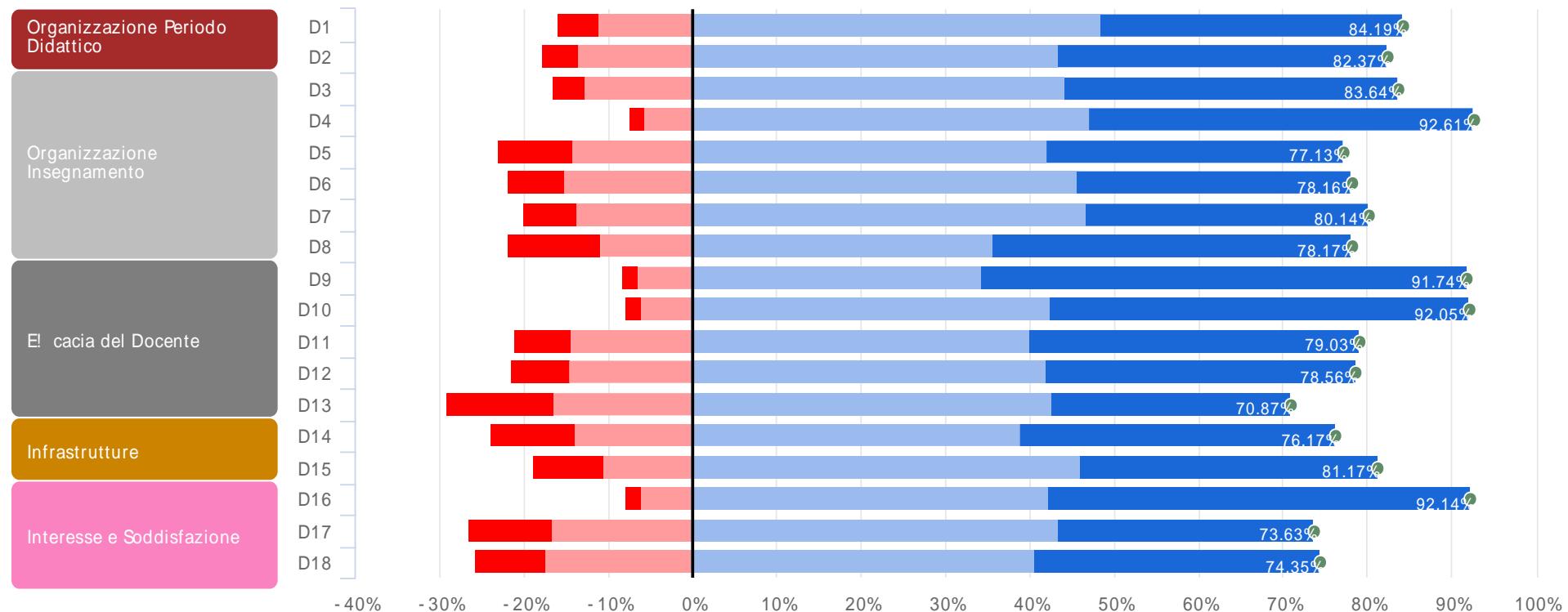
It is physically impossible to modify -6 files

Figure 5. Mean files per changeset.

Likert / Agreement

- Likert scale:
 - ◆ Measures agreement / disagreement with a given statement
 - ◆ Response on an ordinal scale, e.g.
 - Definitely No
 - Mostly No
 - Undecided
 - Mostly Yes
 - Definitely Yes
- Often used to measure positive vs. negative perception

Diverging Stacked Bars

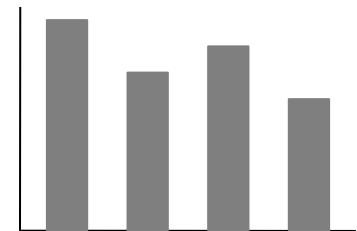
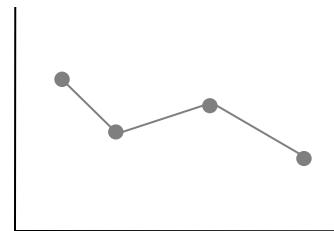
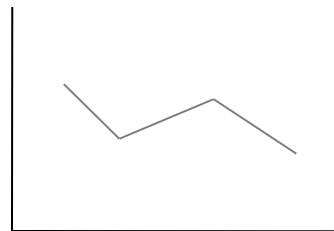


Time series

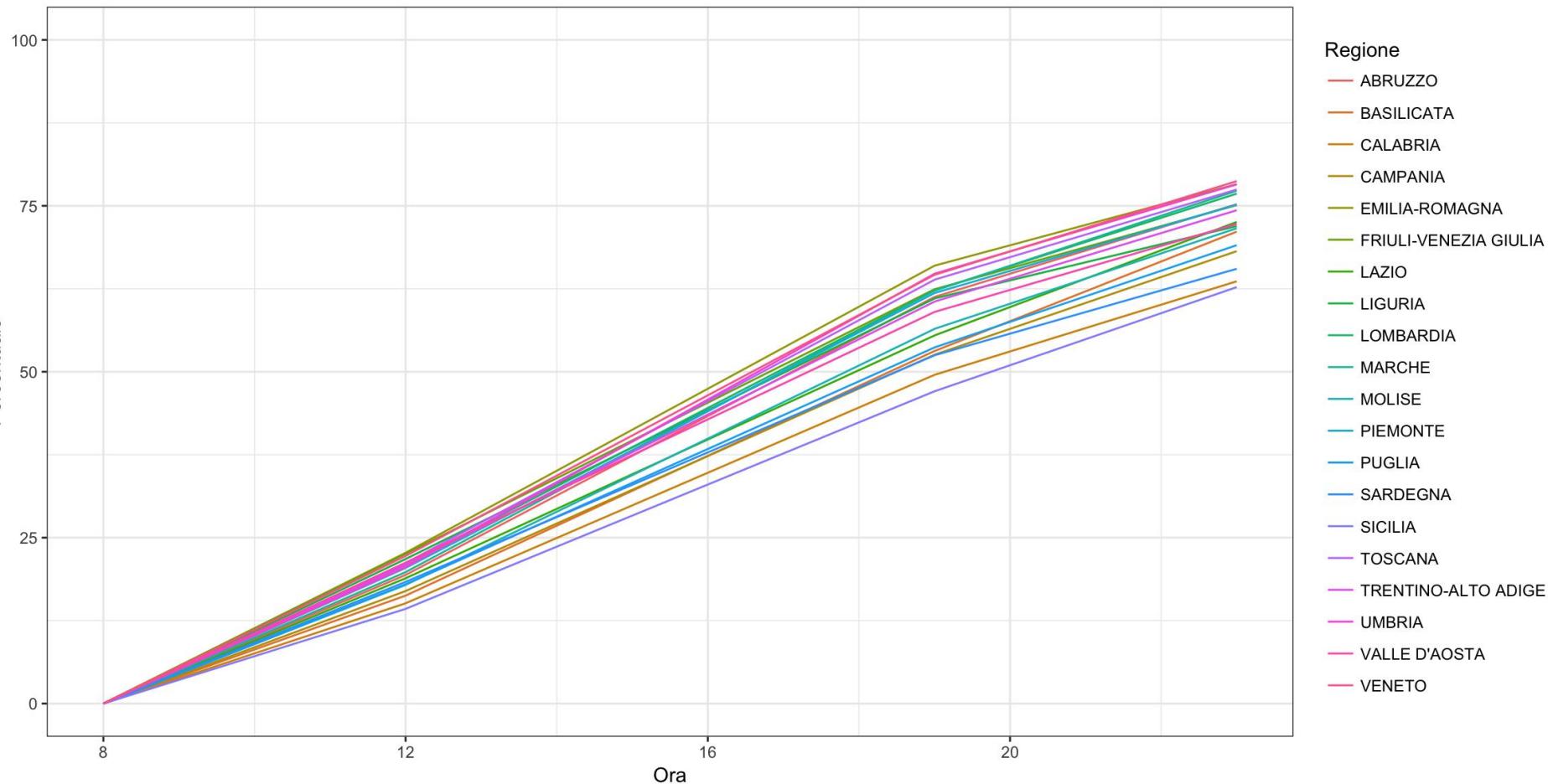
- Series of relationships between quantitative values that are associated with categorical subdivisions of time
- Communicate
 - ◆ Change
 - ◆ Rise
 - ◆ Increase
 - ◆ Fluctuate
 - ◆ Grow
 - ◆ Decline
 - ◆ Decrease
 - ◆ Trend

Time series

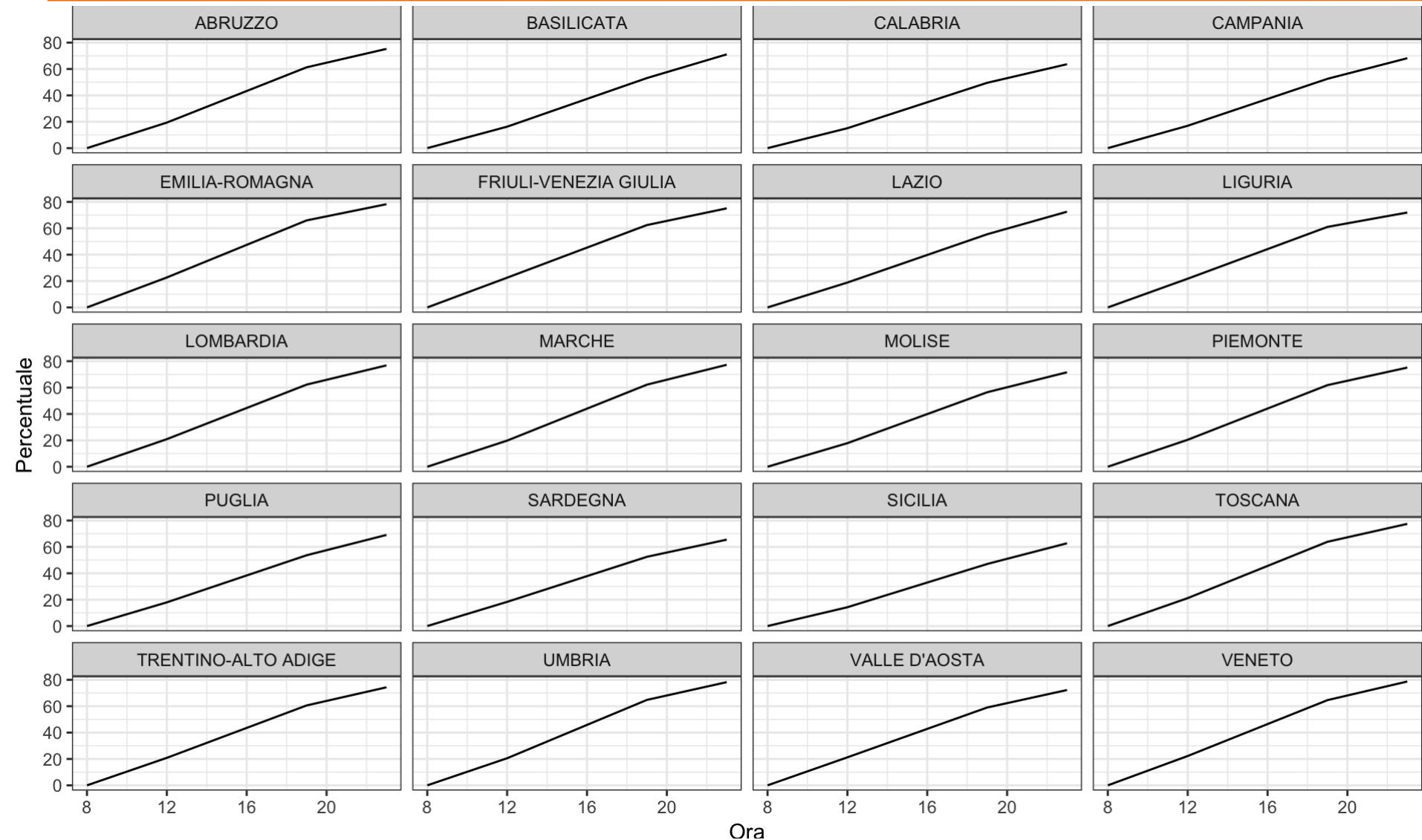
- Time grows from left to right
 - ◆ Cultural convention
- Vertical bars
 - ◆ highlight individual points in time
 - ◆ hide overall trend



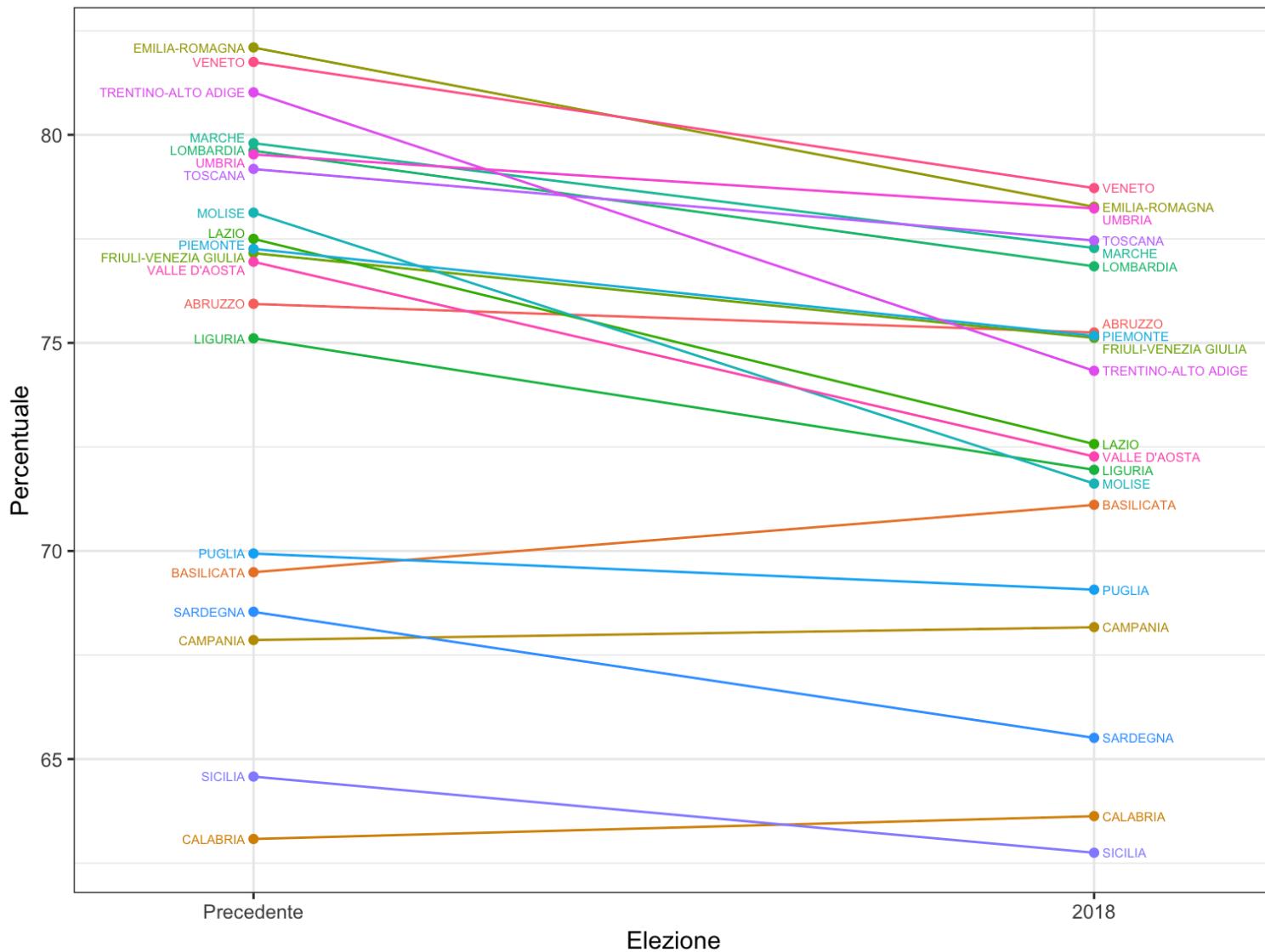
Trends



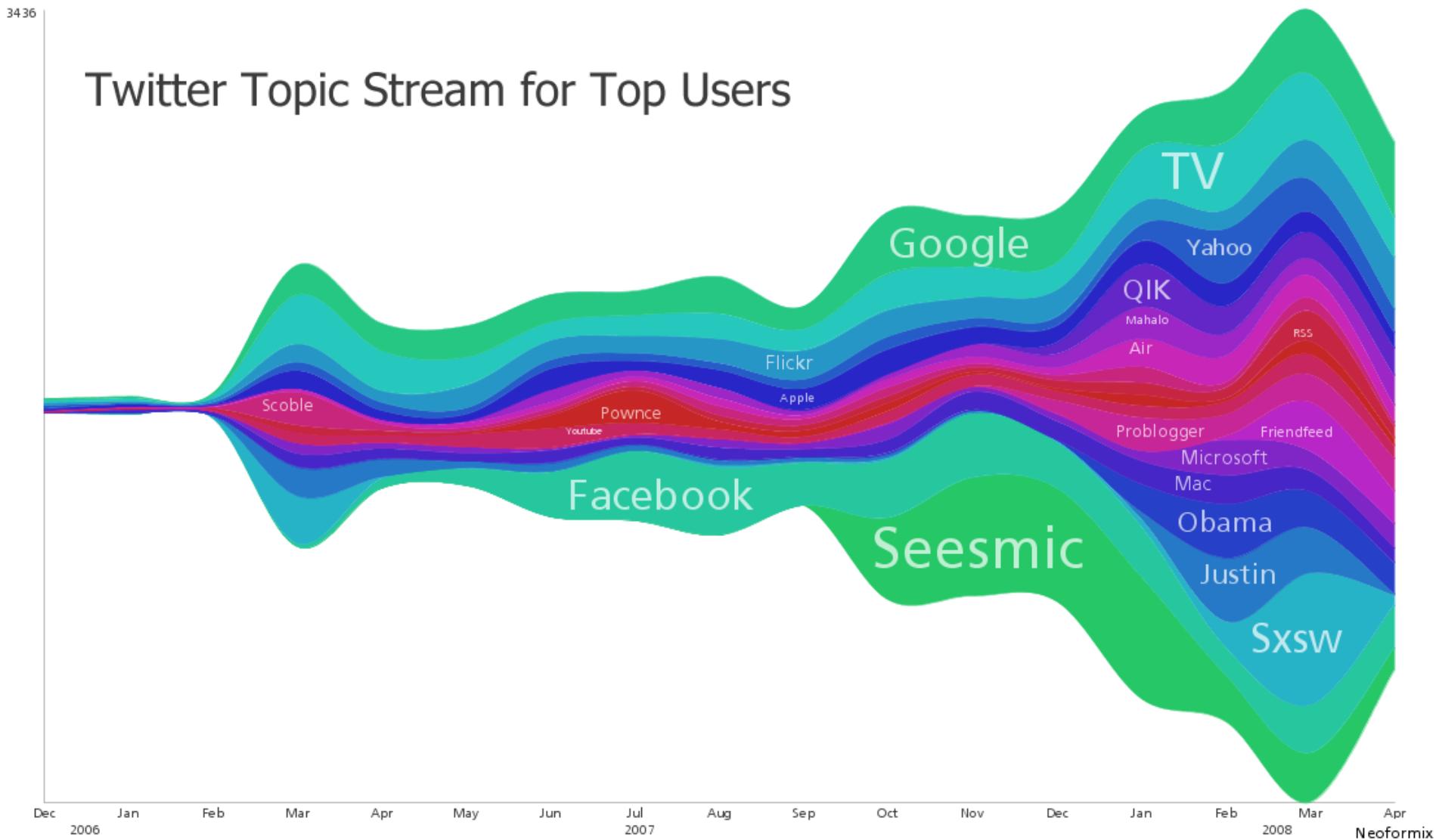
Lines small multiples



Slope chart



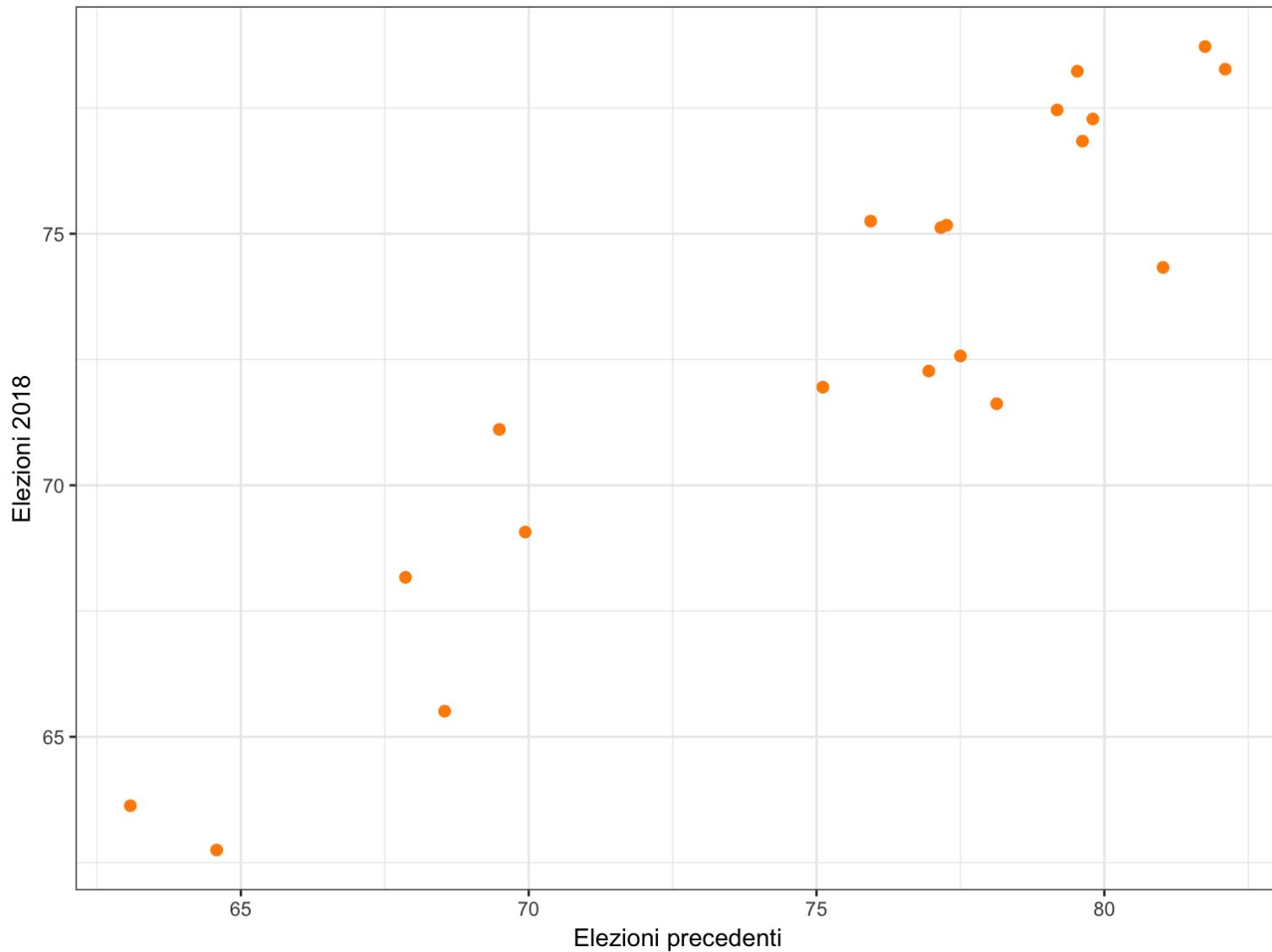
Streamgraph



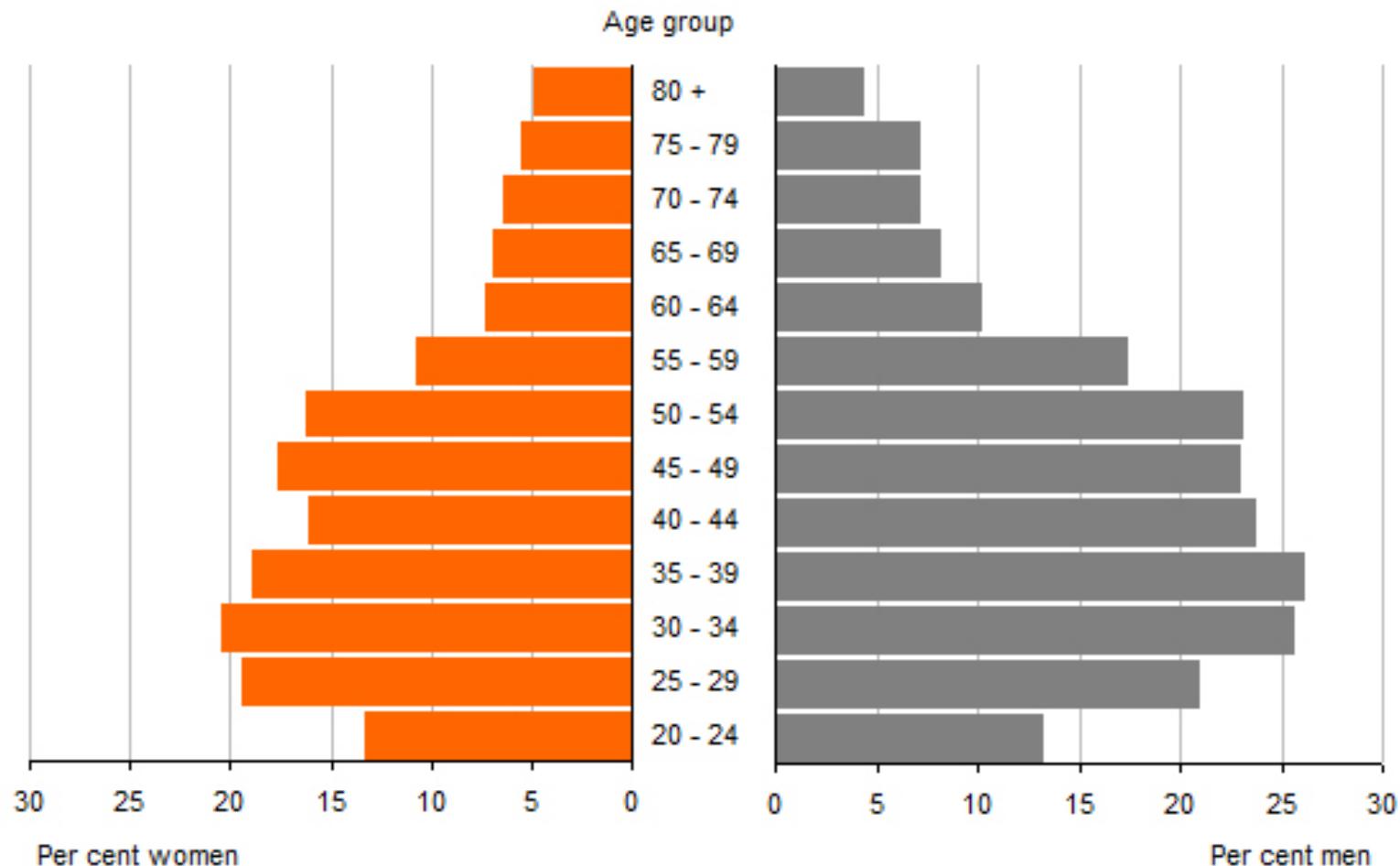
Correlation

- Relationships between two paired sets of quantitative values
 - ◆ Scatter plot w/possible trend line
 - Ok for educated audience
 - ◆ Paired bar graph

Scatter plot



Paired bargraph



[https://unstats.un.org/unsd/genderstatmanual/
Print.aspx?Page=Presentation-of-gender-statistics-in-graphs](https://unstats.un.org/unsd/genderstatmanual/Print.aspx?Page=Presentation-of-gender-statistics-in-graphs)

DASHBOARD

Dashboard

Visualization of the most relevant information

needed to achieve one or more goals
which fits entirely on a single screen
so it can be monitored at a glance

Dashboard

- Dashboards display mechanisms are
 - ◆ small
 - ◆ concise
 - ◆ clear
 - ◆ intuitive
- Dashboards are customized
 - ◆ To suit the goals of person, group, function

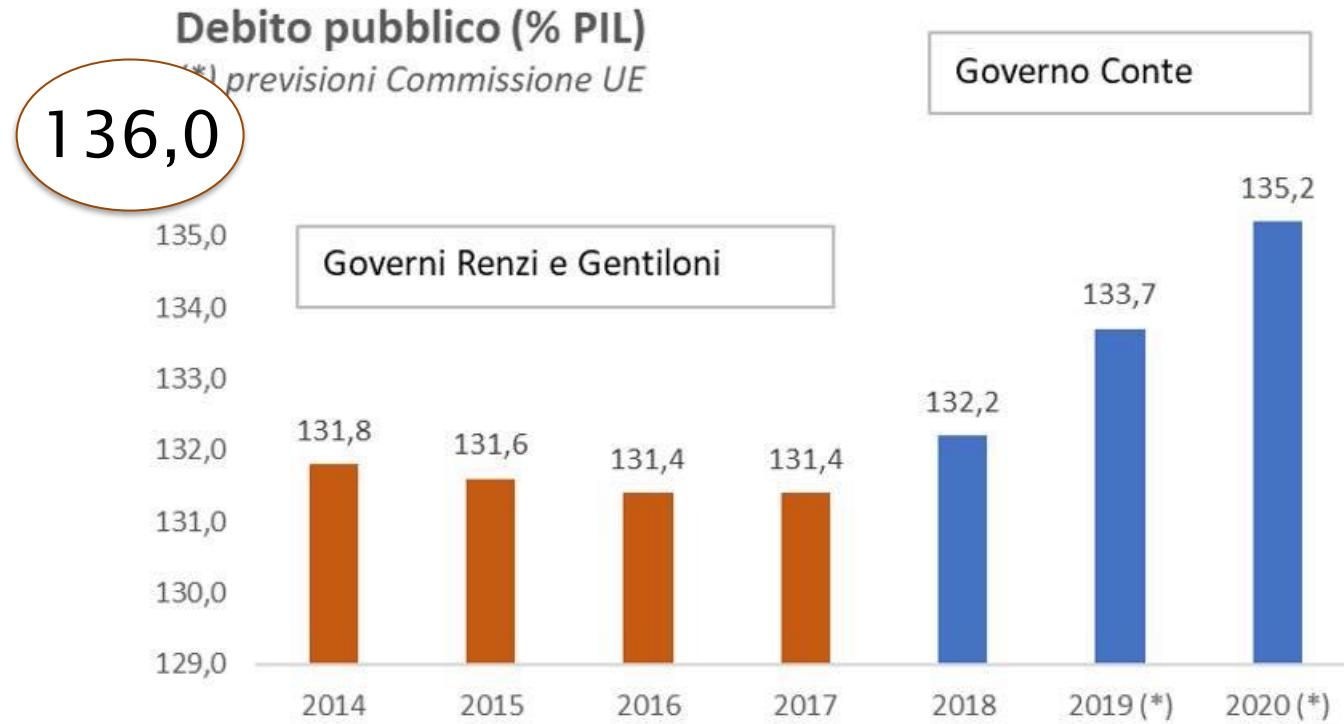
Provide context for data

- References allow judging the data



Use appropriate detail

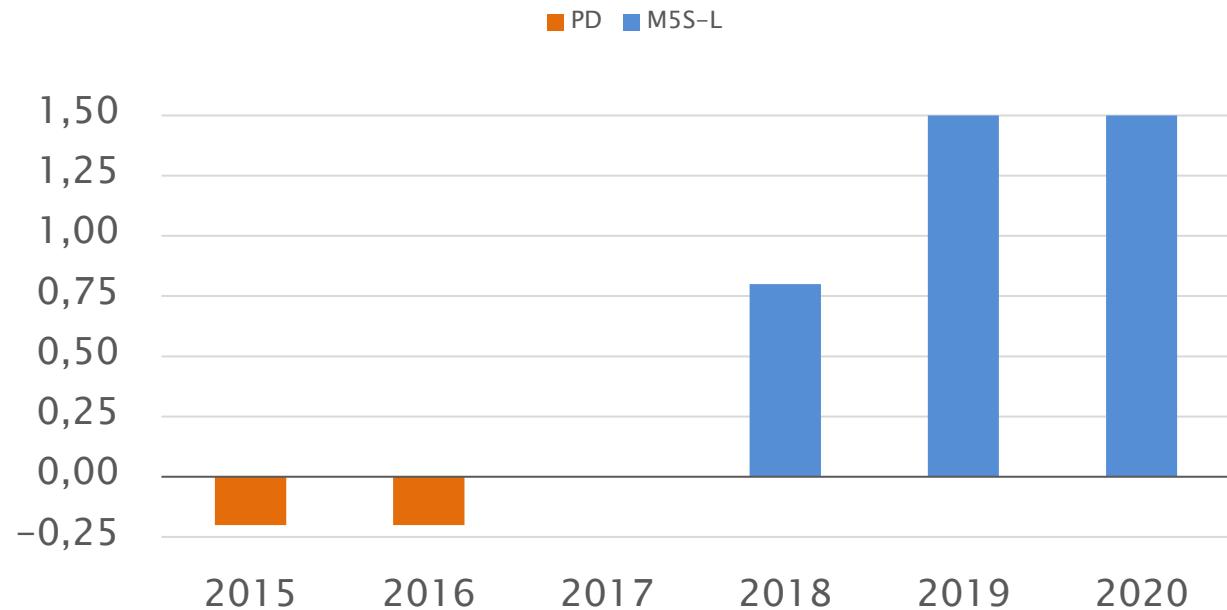
- Typical counter-examples
 - ◆ Dates with seconds detail
 - ◆ Decimals



Use the right measures

- If you are interested in e.g. the difference, ratio, variation show such derived measure

Variazione Debito Pubblico (% PIL)



Use appropriate visualization

- Typical errors:
 - ◆ Any chart when a table would be better
 - ◆ Pie-charts not representing part-whole
 - ◆ Bubble charts

Visualization instruments

- Tables
 - ◆ Textual information
- Graphs
 - ◆ Visual information

Avoid decorations

- Skeumorphic design
- Backgrounds motives
- Color gradients
- Variations not encoding any measure
 - ◆ Typically color

Avoid decorations

- Skeuomorphic design
- Backgrounds motives
- Color gradients
- Variations not encoding any measure
 - ◆ Typically color



-VS-



Avoid decorations

- Skeumorphic design
- Backgrounds motives
- Color gradients
- Variations not encoding any measure
 - ◆ Typically color

A

B

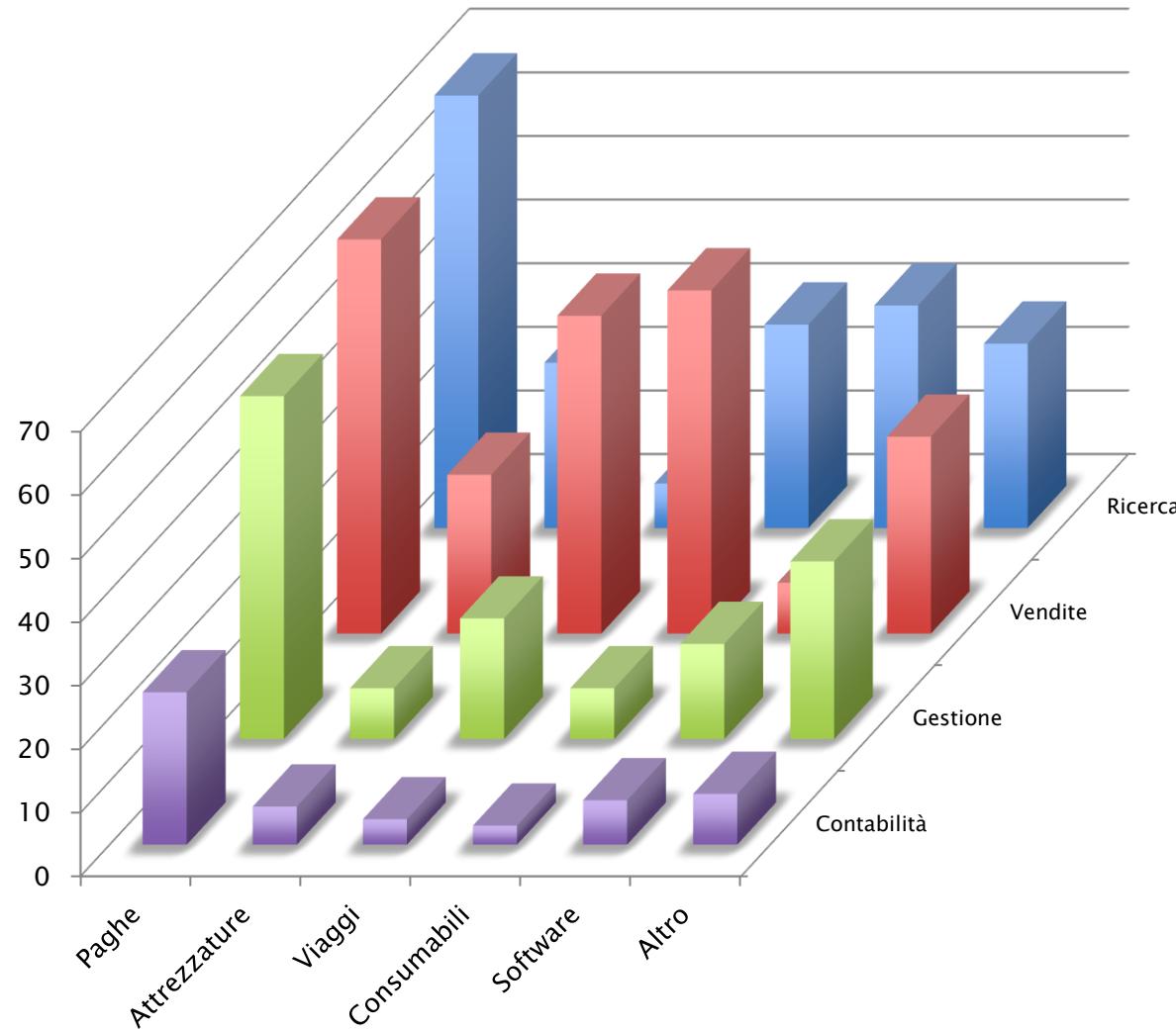
Avoid decorations

- Skeumorphic design
- Backgrounds motives
- Color gradients
- Variations not encoding any measure
 - ◆ Typically color

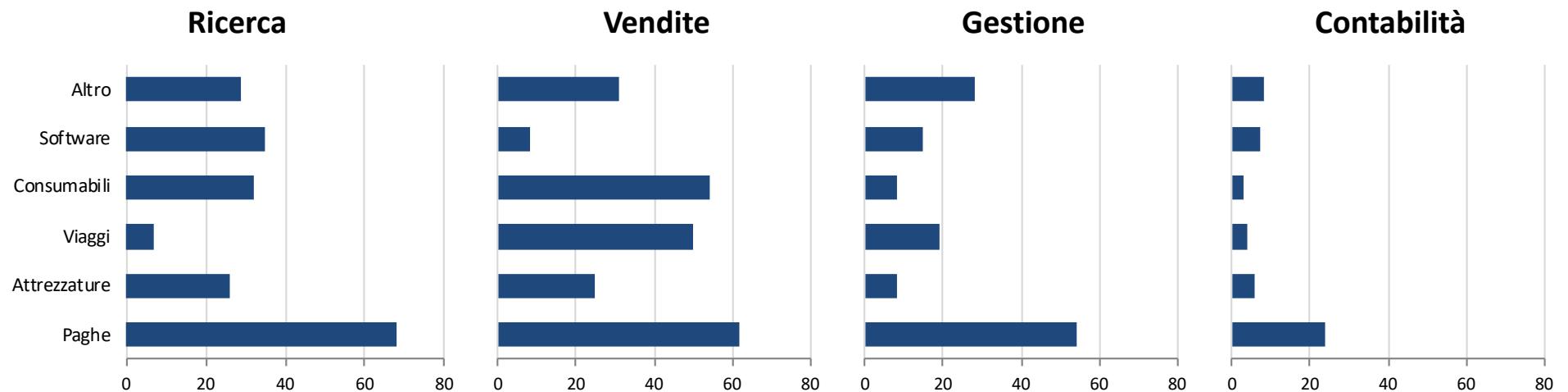
3D diagrams

- Encoding
 - ◆ Axonometry typically hides some data and makes comparison hard
- Not encoding
 - ◆ Perspective deform dimensions
 - ◆ Depth or height distract and make comparison more difficult

Encoding 3D

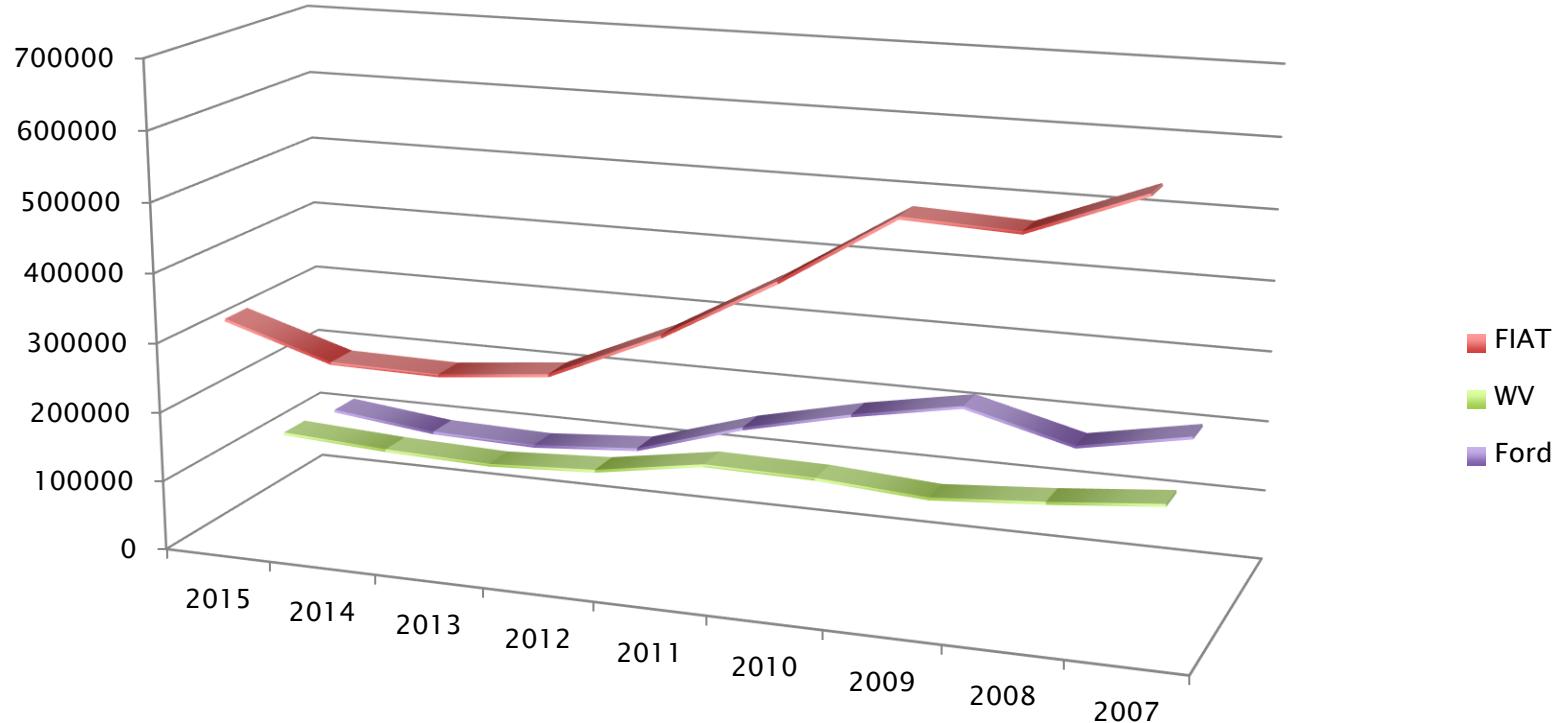


Encoding 3D → 2D



Decorative 3D

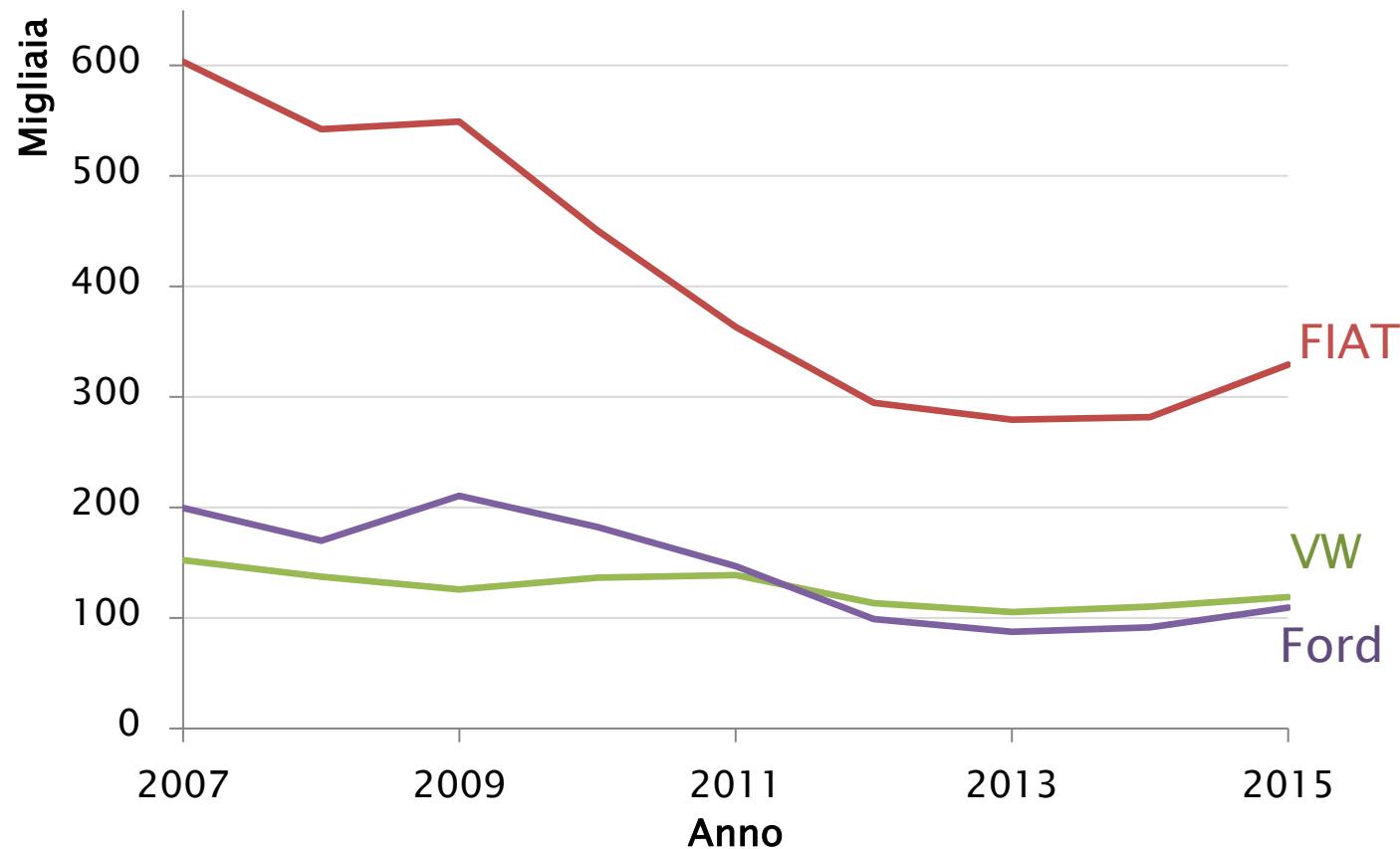
Immatricol.



Decorative 3D → 2D

Immatricolazioni auto per marchio sul mercato italiano

Immatricol.



References

- Stephen Few, 2004. Show me the numbers. Analytics Press.
 - ◆ <http://www.perceptualedge.com/blog/>
- Edward R. Tufte, 1983. The Visual Display of Quantitative Information. Graphics Press.

References

- Wilkinson, L. (2006). *The grammar of graphics*. Springer Science & Business Media.
- Wickham, H. (2010). A layered grammar of graphics. *Journal of Computational and Graphical Statistics*, 19(1), 3–28.
- Visual Vocabulary
<http://ft.com/vocabulary>

References

- R.Olson. Revisiting the vaccine visualization
 - ◆ <http://www.randalolson.com/2016/03/04/revisiting-the-vaccine-visualizations/>
- Nathan Yau. 9 Ways to Visualize Proportions – A Guide
 - ◆ <http://flowingdata.com/2009/11/25/9-ways-to-visualize-proportions-a-guide/>
- M.Correll, and M.Gleicher. Error Bars Considered Harmful: Exploring Alternate Encodings for Mean and Error *IEEE Transactions on Visualization and Computer Graphics*, Dec. 2014
 - ◆ <http://graphics.cs.wisc.edu/Papers/2014/CG14/Preprint.pdf>

Data Quality

Data Management and Visualization



SoftEng
<http://softeng.polito.it>

Version 1.2.2
© Marco Torchiano, 2021



Licensing Note



This work is licensed under the Creative Commons Attribution–NonCommercial–NoDerivatives 4.0 International License.

To view a copy of this license, visit

<http://creativecommons.org/licenses/by-nc-nd/4.0/>.

You are free: to copy, distribute, display, and perform the work

Under the following conditions:

 **Attribution.** You must attribute the work in the manner specified by the author or licensor.

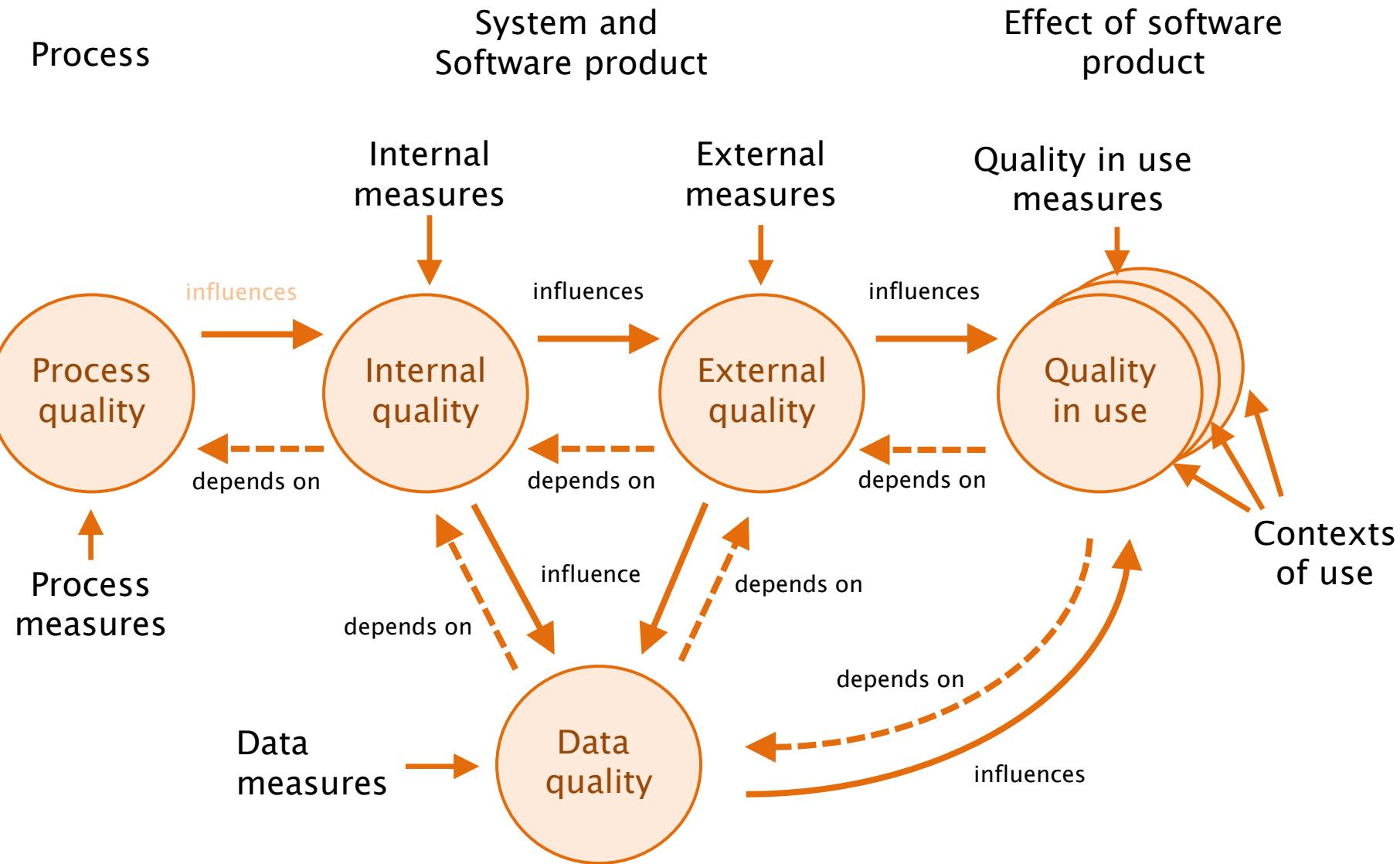
 **Non-commercial.** You may not use this work for commercial purposes.

 **No Derivative Works.** You may not alter, transform, or build upon this work.

- For any reuse or distribution, you must make clear to others the license terms of this work.
- Any of these conditions can be waived if you get permission from the copyright holder.

Your fair use and other rights are in no way affected by the above.

Software Qualities



Target entities

Information System

ICT Product

Data

Software

Component

Hw & Communication

Context of use

Users

Goals

User Environ.

System Context

Target entities vs. Q. Models

Data Quality

Product Quality

Quality in use

Information System

ICT Product

Data

Software

Component

Hw & Communication

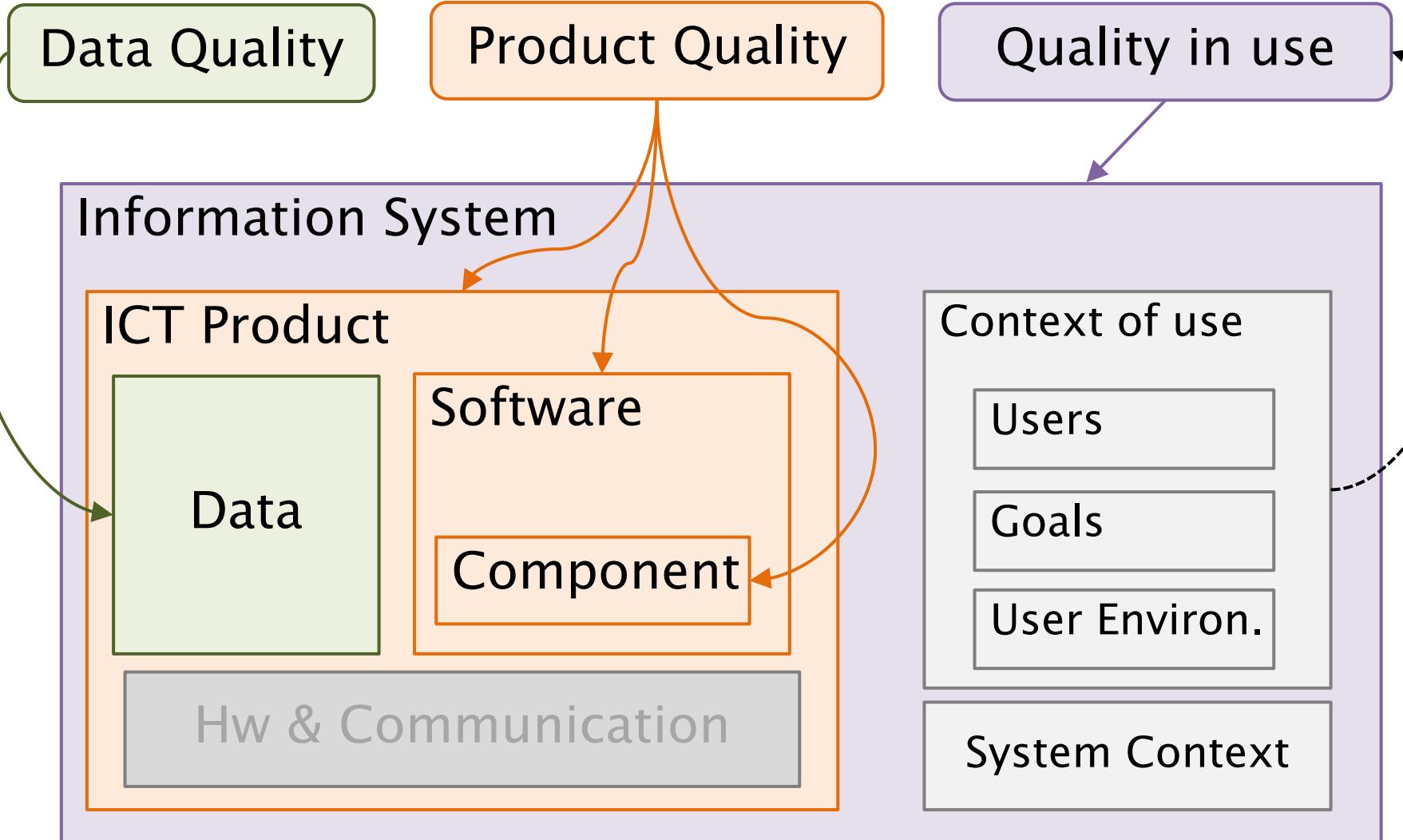
Context of use

Users

Goals

User Environ.

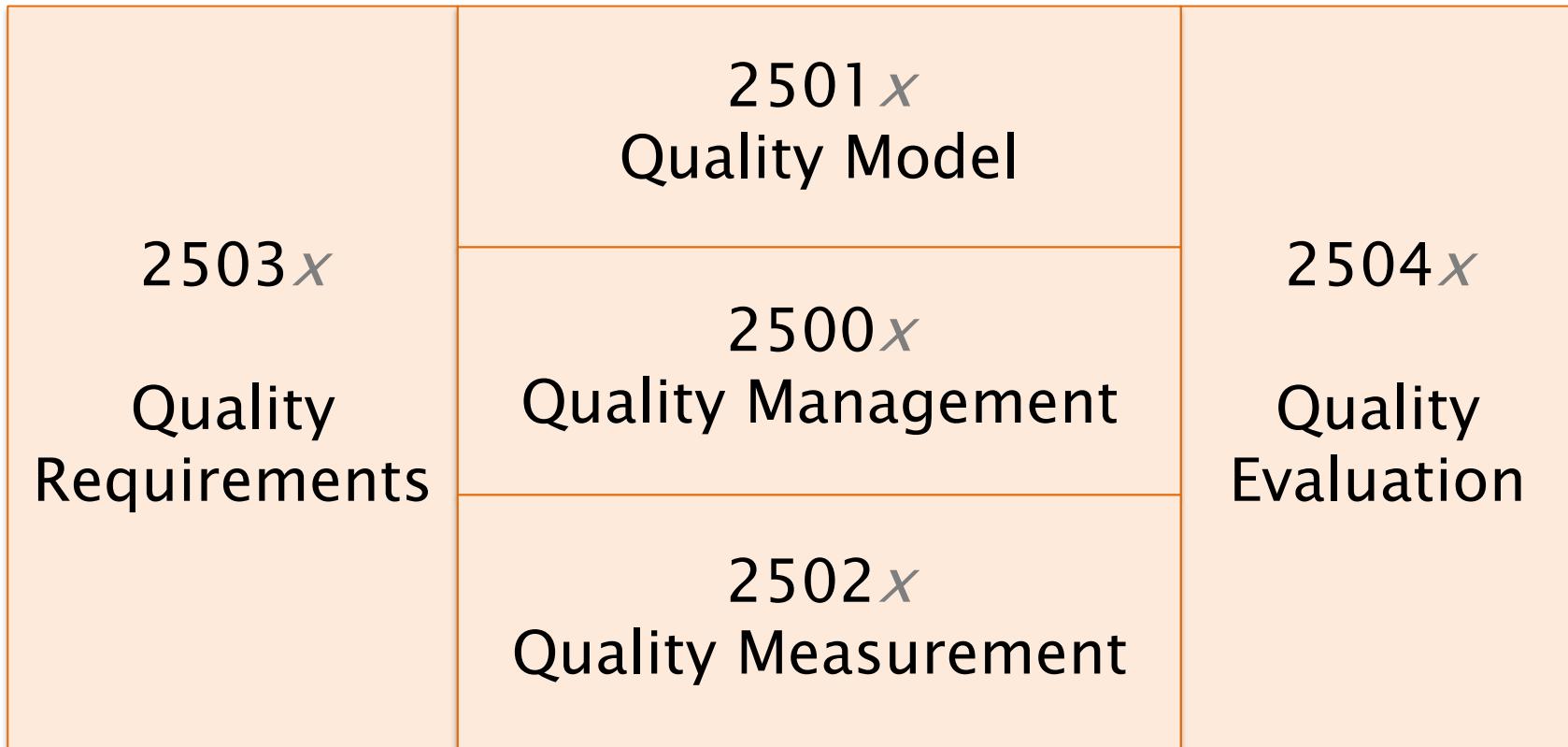
System Context



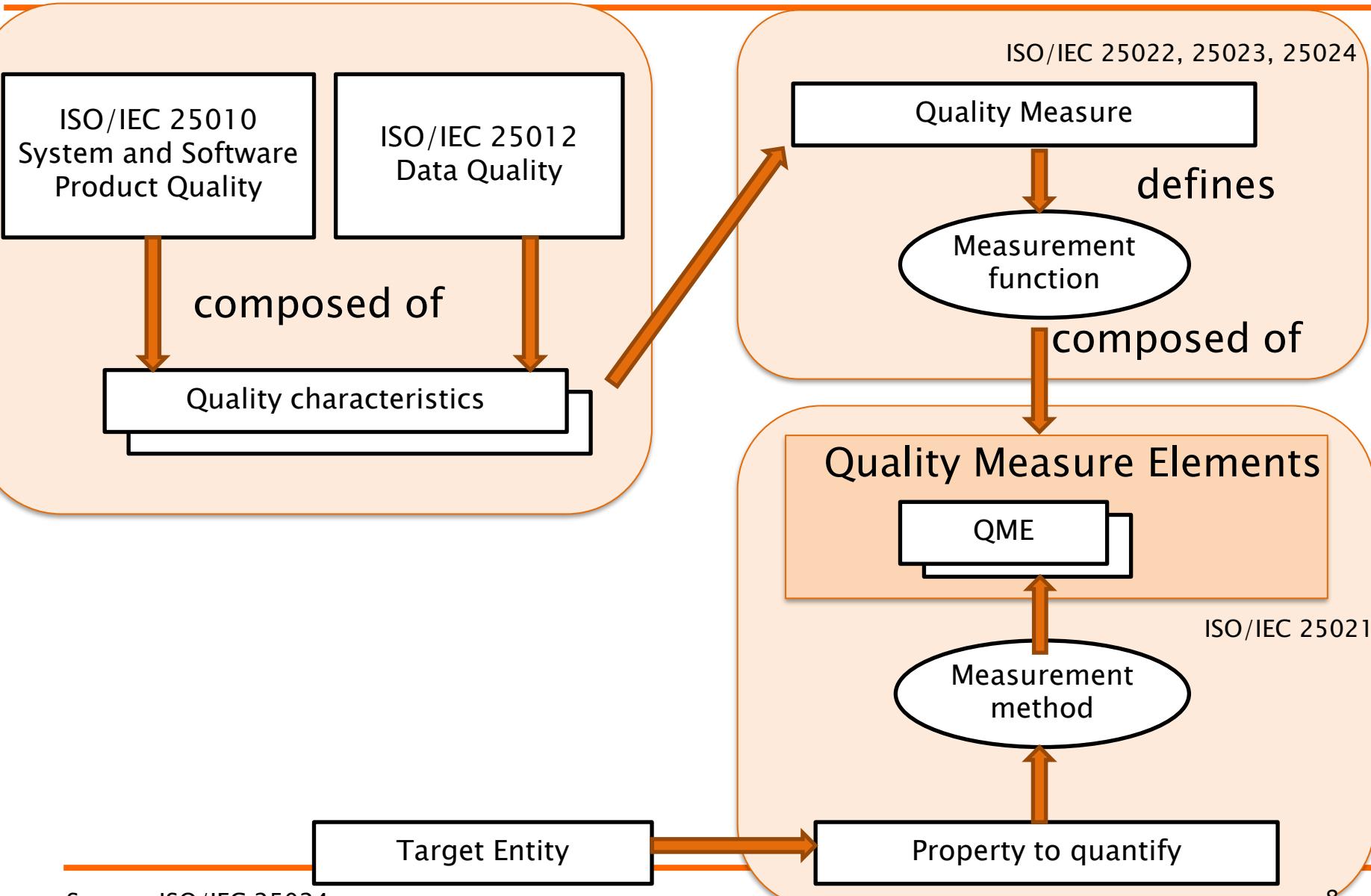
Software Product Quality

- ISO/IEC 9126: Issued 1991, revised 2001
 - Being retired
- ISO/IEC 250xx – SQuaRE
 - ◆ Software product Quality Requirements and Evaluation
 - ◆ Family of standards
 - in development

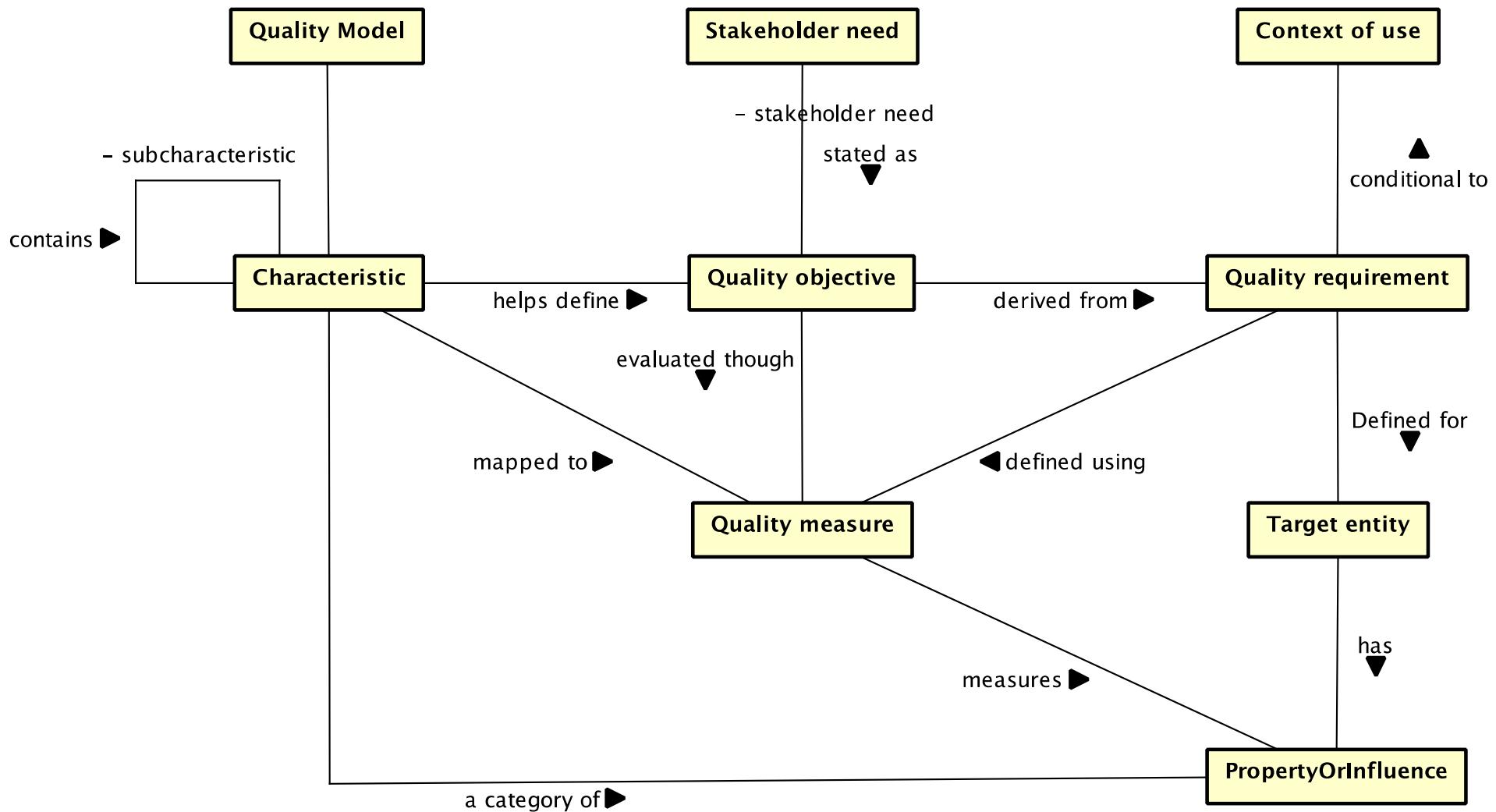
ISO SQuaRE – Standard Family



Relationships among standards



Quality conceptual model



Model structure

- Characteristic
 - ◆ Main aspects, e.g., usability
- Sub-Characteristic
 - ◆ Specific aspects, e.g. accessibility
- Measure
 - ◆ Measurement function to evaluate a specific (sub)-characteristic
- Measure element
 - ◆ Fundamental

DATA QUALITY

Quality characteristics

Inherent: facts

- Accuracy
- Completeness
- Consistency

- Currency
- Credibility

- Accessibility
- Compliance
- Confidentiality
- Efficiency

- Understandability
- Precision
- Traceability

- Availability
- Portability

- Recoverability



System dependent: artefacts

Quality characteristics

- Accuracy
 - Completeness
 - Consistency
-
- Accessibility
 - Compliance
 - Confidentiality
 - Efficiency
-
- Availability
 - Portability
-
- Currency
 - Credibility
-
- Understandability
 - Precision
 - Traceability
-
- Recoverability

Accuracy

- Correspondence between data and reality
 - ◆ Syntactic
 - It belongs to a set of validated information
 - ◆ Semantic
 - The meaning (the content) corresponds to the reality

Open or Closed World?

- **Closed World (CWA):**
 - ◆ The knowledge represented in the data (and its schema) is complete
 - ◆ E.g., if a code appears in the list of valid codes it is correct, otherwise it is wrong
- **Open World (OWA):**
 - ◆ The knowledge represented in the data is (knowingly) incomplete
 - ◆ E.g., if a code appears in the list of valid codes it is correct, otherwise it is not possible to tell for sure

CWA – Accuracy: Genomics

- Human genes are known and coded, each has a predefined symbol
- Any code not included in those predefined represents a syntactic accuracy error
- E.g. code ‘**SEPT2**’(Septin-2) when imported into  is automatically turned into ‘September 2’

OWA – Accuracy

How to decide what is accurate?

- Rules that define what is syntactically correct
 - ◆ E.g. regular expressions
- Constraints to define what values are semantically acceptable
 - ◆ E.g. validity interval

Where do rules come from?

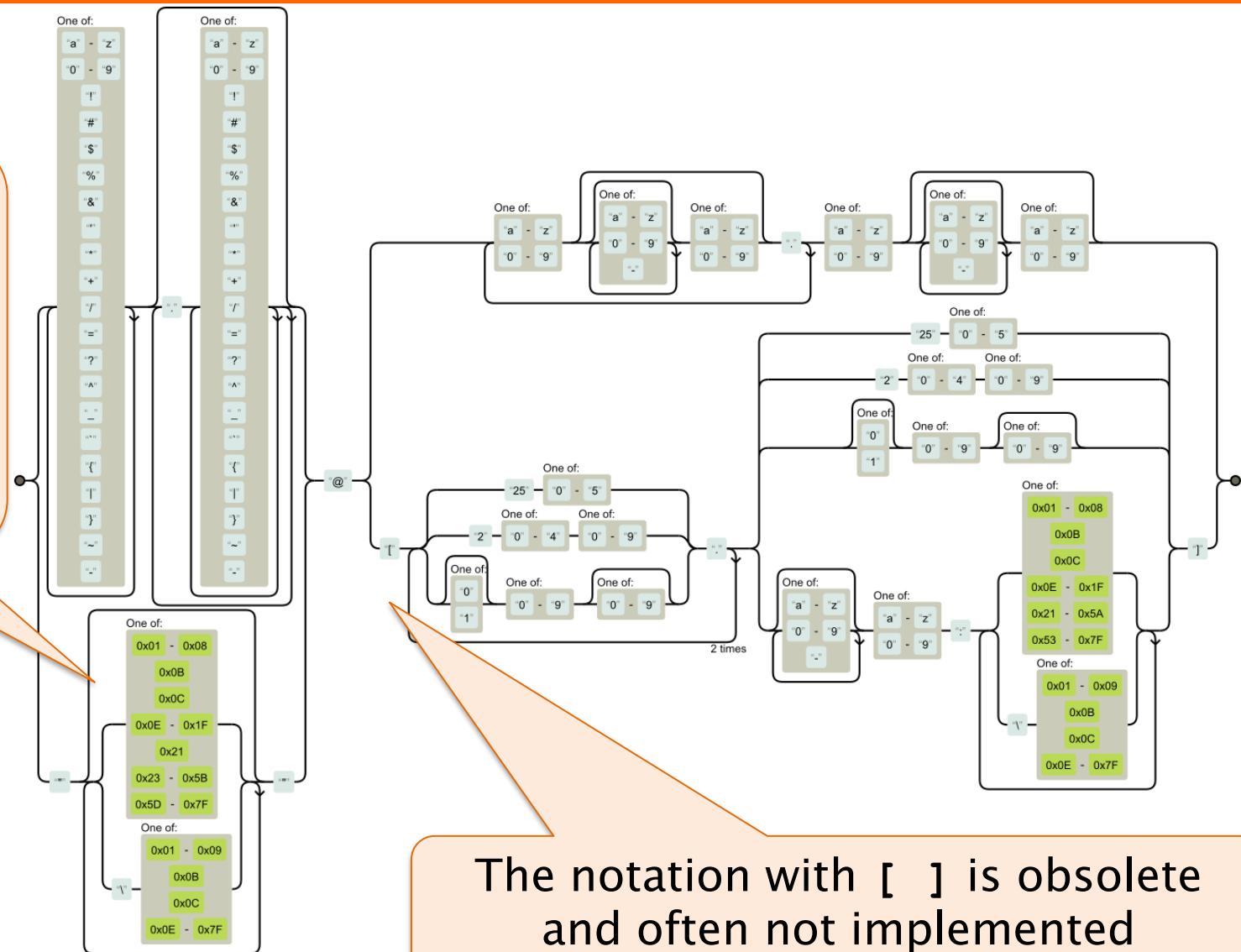
- Standard
- Domain knowledge
- Similar data
- Past data

OWA: Email per RFC-5322

```
\A(?:[a-zA-Z!#$%&'*/=?^`{|}~-]+(?:\.\[a-zA-Z!#$%&'*/=?^`{|}~-]+)*  
| "(?:[\x01-\x08\x0b\x0c\x0e-\x1f\x21\x23-\x5b\x5d-\x7f]  
| \\\\[ \x01-\x09\x0b\x0c\x0e-\x7f])*")  
@ (?:(?:[a-zA-Z] (?:[a-zA-Z-]*[a-zA-Z])?\.\.)+[a-zA-Z] (?:[a-zA-Z-]*[a-zA-Z])?  
| \[(?:25[0-5]|2[0-4][0-9]| [01]?[0-9][0-9]?)\.\.){3}  
(?:25[0-5]|2[0-4][0-9]| [01]?[0-9][0-9]?|[a-zA-Z-]*[a-zA-Z]:  
|(?:[\x01-\x08\x0b\x0c\x0e-\x1f\x21-\x5a\x53-\x7f]  
| \\\\[ \x01-\x09\x0b\x0c\x0e-\x7f]))+)  
\])\z
```

OWA: Email per RFC-5322

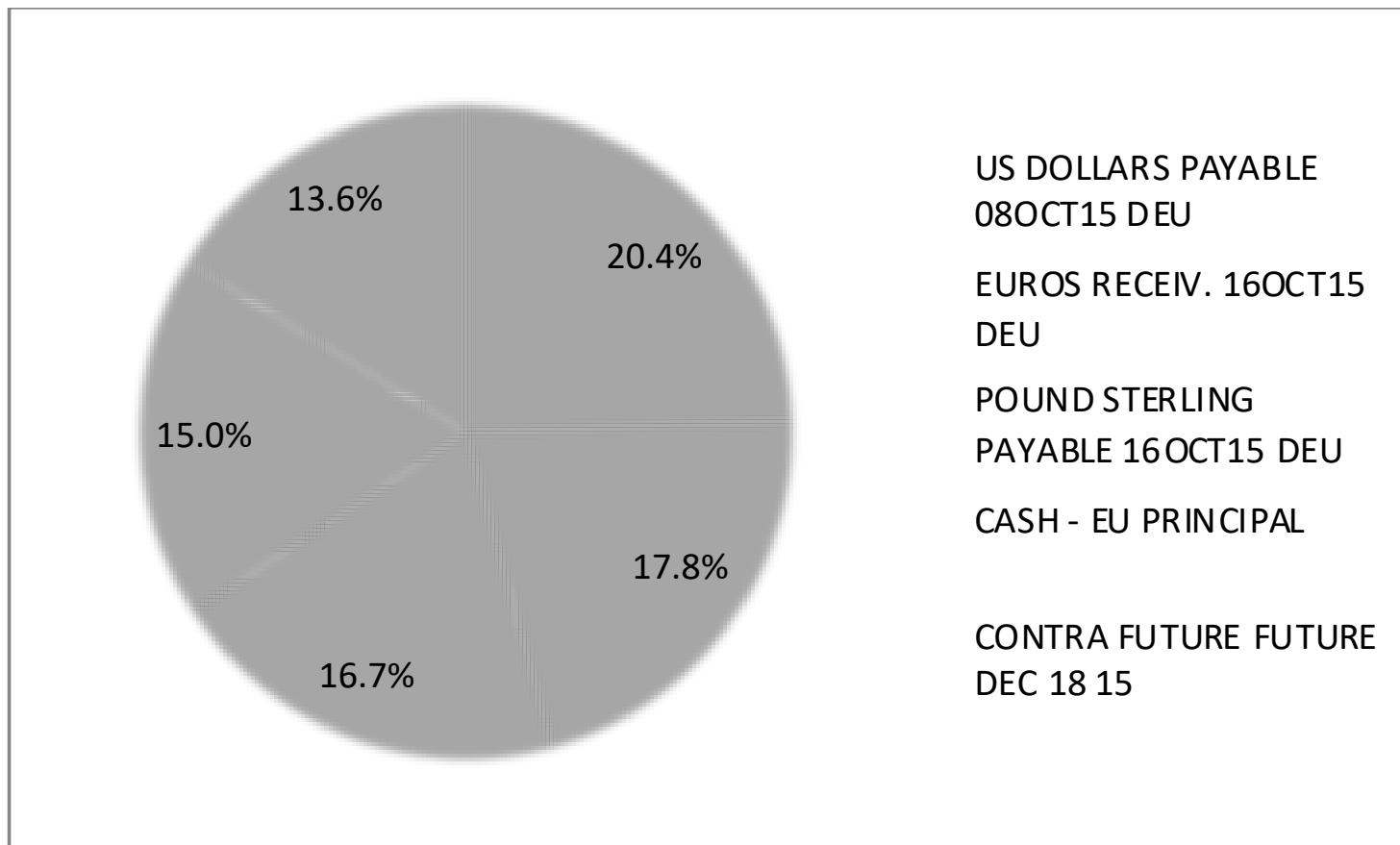
Non
printable
characters
are usually a
problem for
email clients



Completeness

- Computer: presence of all necessary values
 - ◆ Both to entity occurrences and to attributes of a single occurrence
 - ◆ Note: not all missing values constitute a completeness issue
 - User: how much the available data is capable of satisfying the needs
-

Completeness



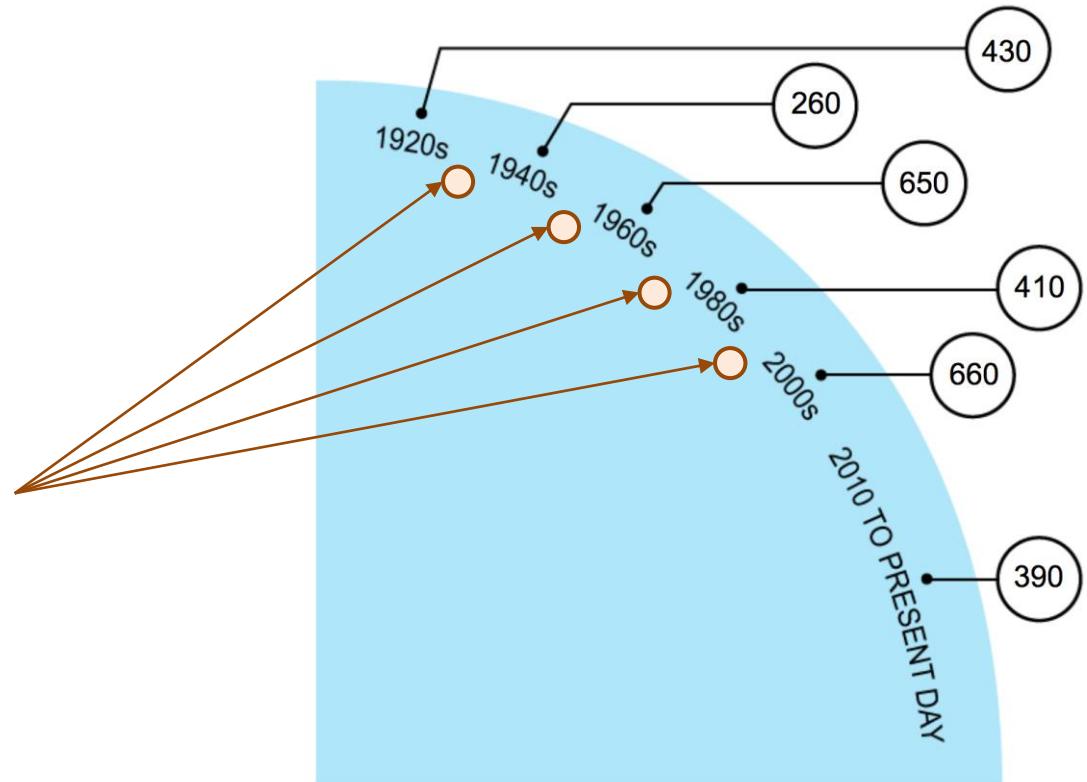
Sum of percentages: 83.5%
We miss the remaining 16.5%

Also consistency:
expected 100%

Completeness

REINVENTING THE WIPER

Number of windshield-wiper-related patents issued per decade.



What about
1930s, 1950s,
1970s, 1990s ?

A possible hypothesis,
another one considered later

Consistency

- Absence of contradictions in the data
 - ◆ Referential integrity
 - Often guaranteed in RDBMS
 - ◆ Duplication
 - Increase the risk of inconsistency on update
 - ◆ Semantic
 - E.g. birth date must be before death date

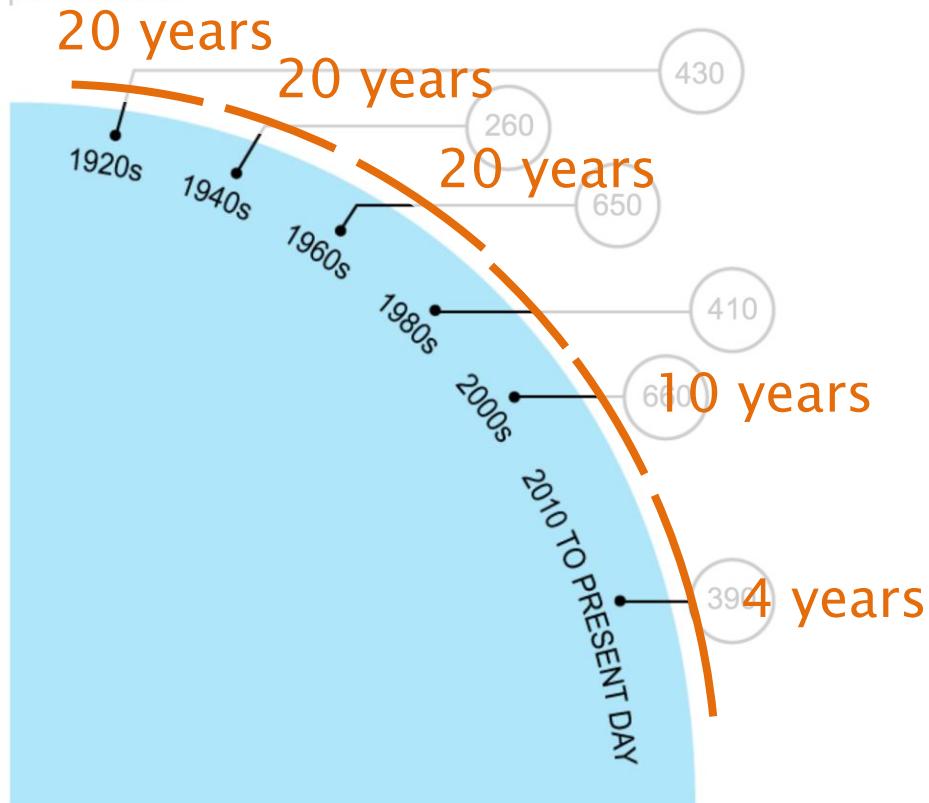
Consistency in graph data

- Values in a series of data encoded with visual attributes must be comparable
 - ◆ Consistent aggregation level
 - ◆ Consistent measurement method
 - ◆ Consistent target entities

Aggregation level

REINVENTING THE WIPER

Number of windshield-wiper-related patents issued per decade.



Count on of events on periods of different length are not comparable

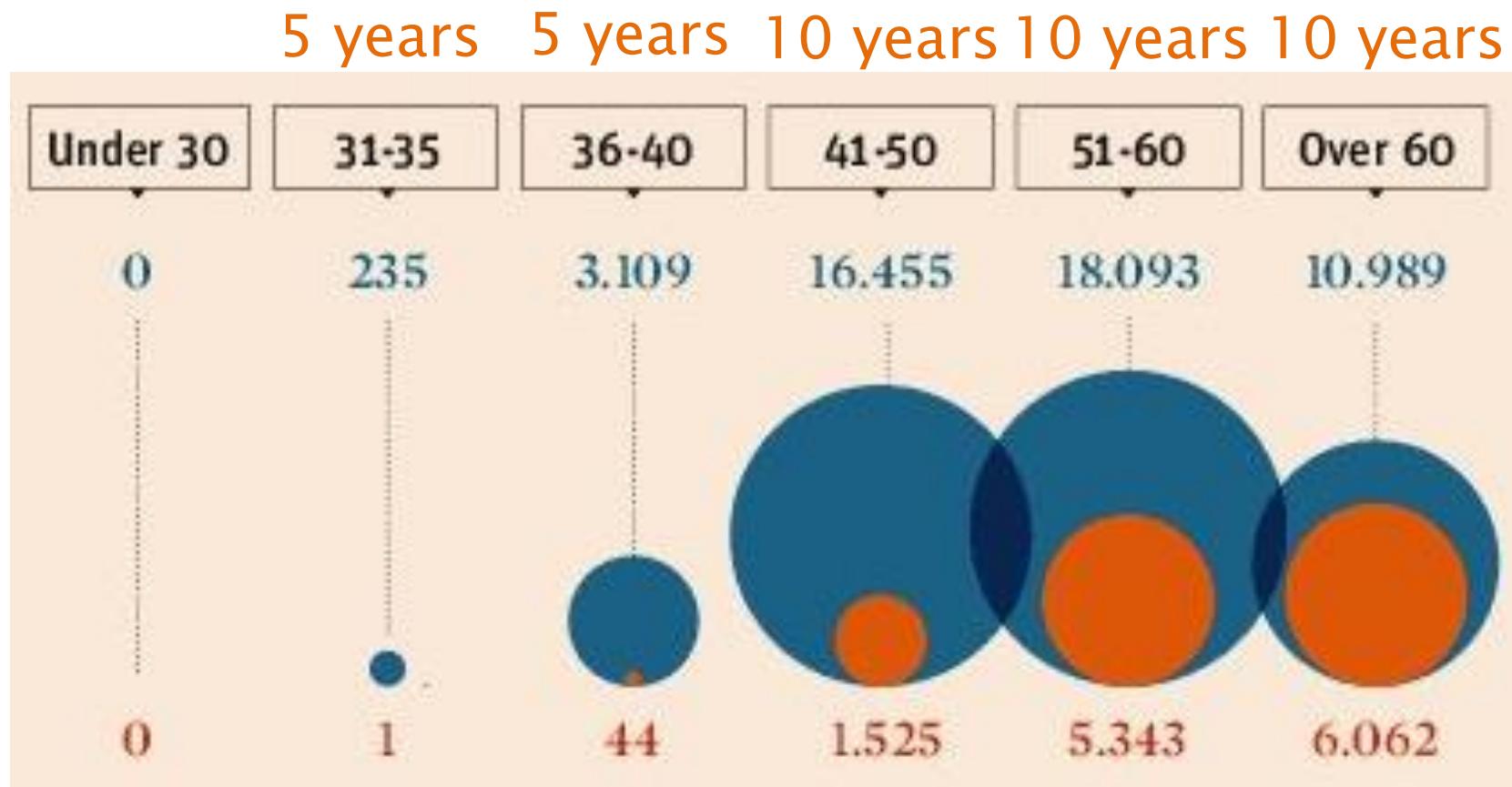
A possible hypothesis,
another one considered earlier

Aggregation level

Period	Duration [years]	Patents	Pat. per year
1920s	20	430	21.5
1940s	20	260	13.0
1960s	20	650	32.5
1980s	20	410	20.5
2000s	10	660	66.0
2010 to present	4	390	97.5

When comparing values corresponding to entities or categories with different *size*, normalized values (i.e. densities) are comparable, absolute values are not!

Aggregation level



Aggregation level

Range	Size	Count	Density
31–35	5	235	47.0
36–40	5	3109	621.8
41–50	10	16455	1645.5
51–60	10	18093	1809.3
Over 60	10	10989	1098.9

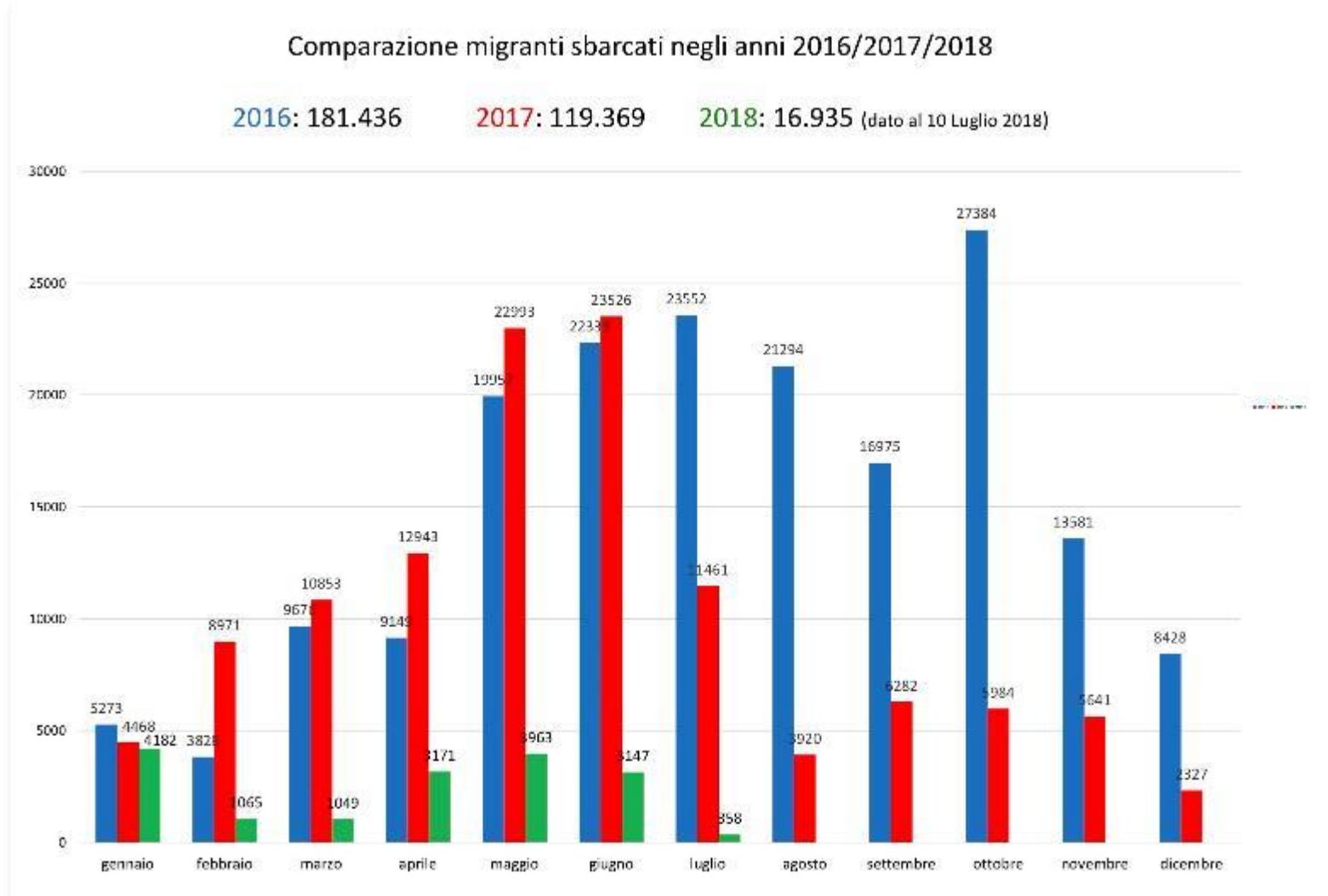
Ratios:

5.3

2.6

Lie factor = 2

Consistent timeframe



Fonte: Dipartimento della Pubblica sicurezza

Consistent timeframe

Year	Months	Value	Normalized
2016	12.0	181 436	15119.7
2017	12.0	119 369	9947.4
2018	6.3	16 935	2688.1

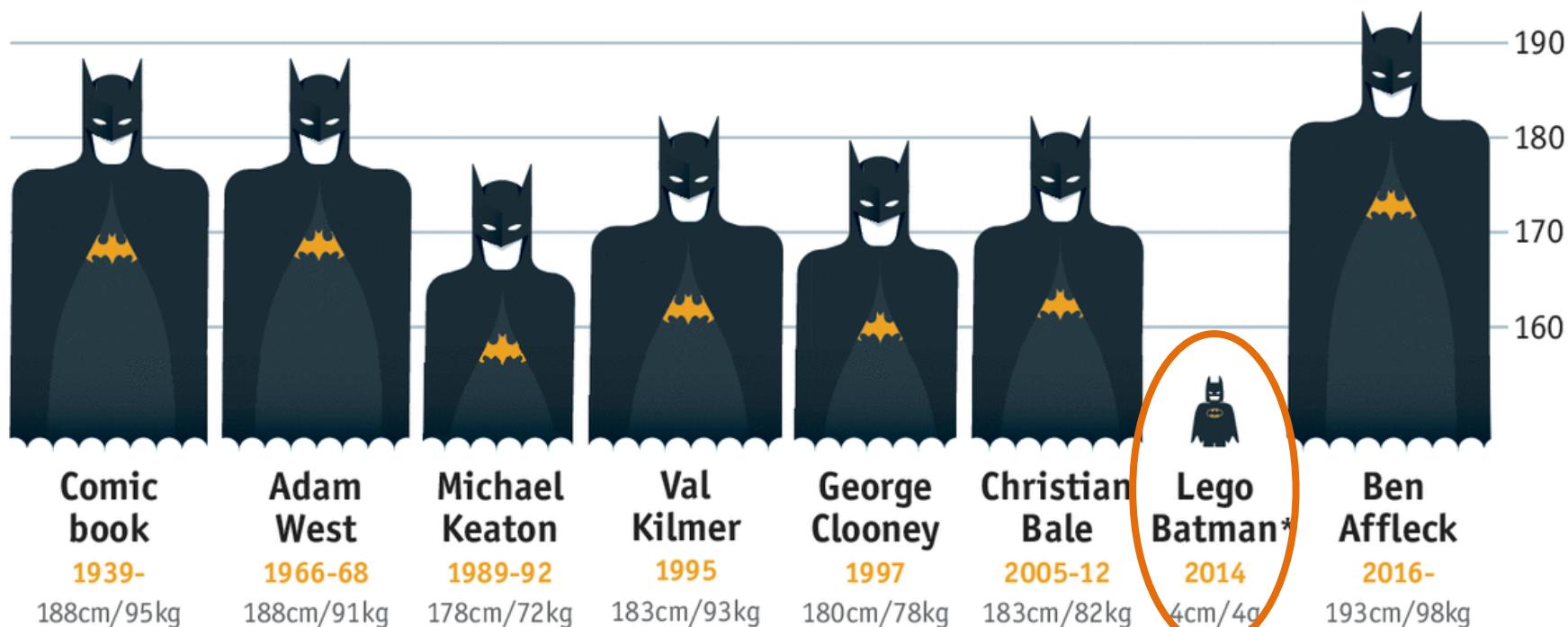
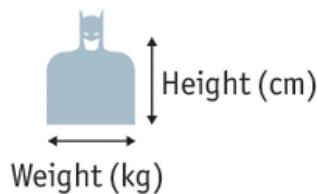
Ratios: 7.0 3.7

Lie factor = 1.9

Consistent target entities

Bruce gain

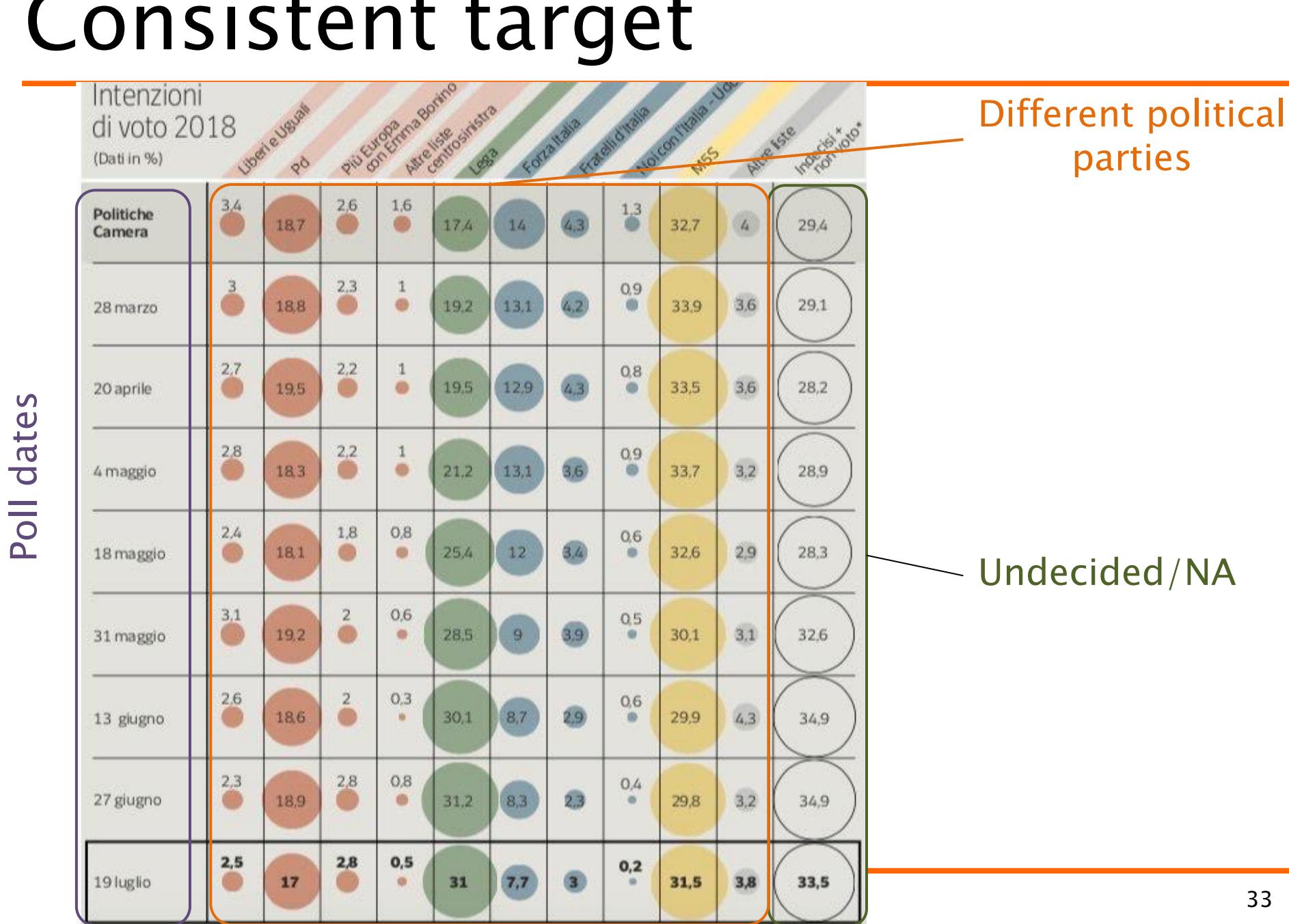
Estimated heights and weights
of on-screen Batmen



Sources: Moviepilot; IMDb

* From "The Lego Movie", not to scale

Consistent target



Consistent target

- Proportions computed on different reference wholes

$$Undecided = \frac{n_{undec} + n_{NA}}{N_{sample}}$$

$$P_i = \frac{n_{pi}}{N_{sample} - n_{undec} - n_{NA}}$$

Consistent method

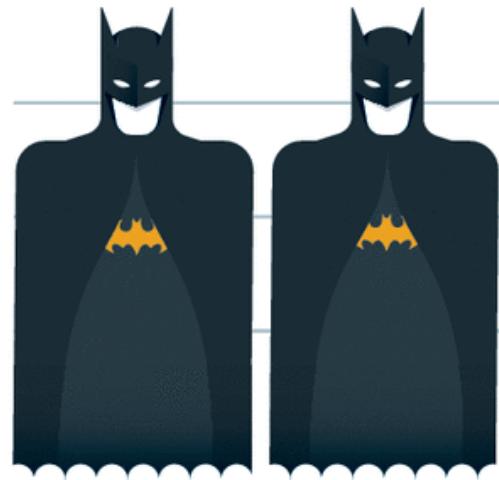
- A series of values that are not measured using the same method **might** not be directly comparable
 - ◆ estimate vs. actual, projection vs. final
 - ◆ periodic samples collected at different possibly non-equivalent times
 - e.g. different period of year, week, day

Currency

- Currency is the extent to which data is up-to-date
 - ◆ With reference to the reality and
 - ◆ With reference to the task at hand
 - Lack of information to establish currency is an Understandability issue
-

Credibility

- The extent to which data are regarded as true and credible by users
- What is the source of the data showed in the graph?



Comic
book
1939-
188cm/95kg

Adam
West
1966-68
188cm/91kg

Sources: Moviepilot; IMDb

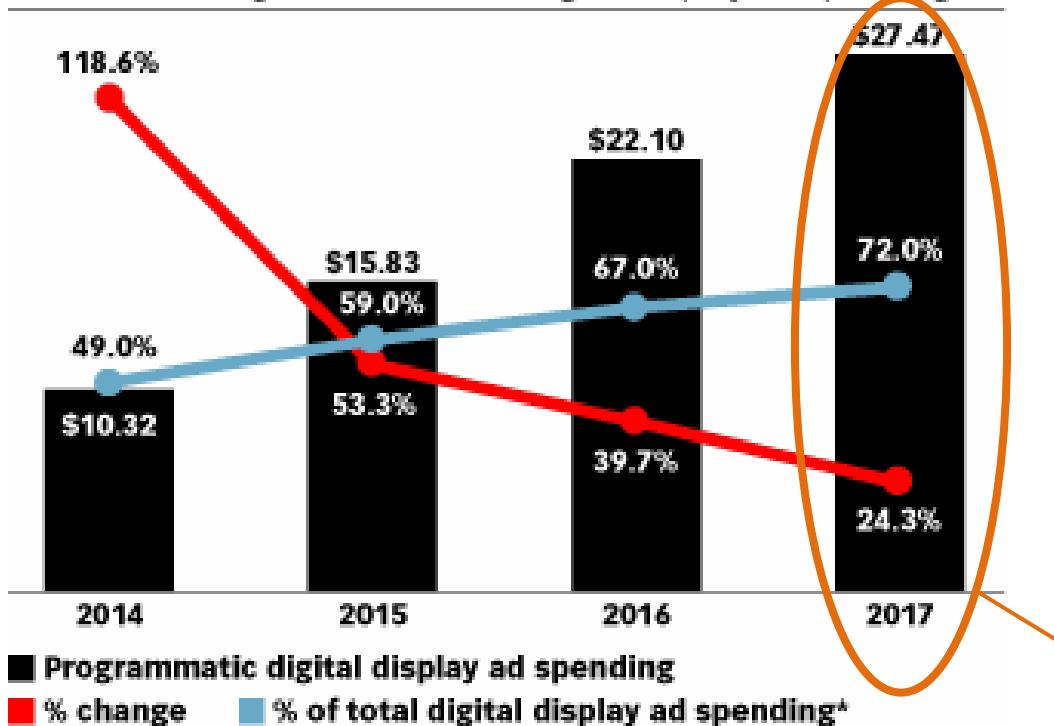
Understandability

- The extent to which data can be read and interpreted by users
- How is data measured? Is there a track of how values are collected, measured or estimated?
 - ◆ If multiple methods are used that might represent an inconsistency issue.

Understandability

US Programmatic Digital Display Ad Spending, 2014-2017

billions, % change and % of total digital display ad spending*



Note: digital display ads transacted via an API, including everything from publisher-erected APIs to more standardized RTB technology; includes native ads and ads on social networks like Facebook and Twitter; includes advertising that appears on desktop/laptop computers, mobile phones, tablets and other internet-connected devices; *includes banners, rich media, sponsorship, video and other

Source: eMarketer April 2016

Data from 2016 including values for 2017.
Undeclared mix of projections and final data.

Precision

- The capability to provide the degree of information needed in a stated context of use
 - ◆ Enough information to allow discriminate
 - ◆ Not too much to overload reader
 - Related to "Utility"

Precision



Precision

Debito pubblico (% PIL)

(*) previsioni Commissione UE

Governo Conte

136,0

135,0

Governi Renzi e Gentiloni

134,0

133,0

132,0

131,0

130,0

129,0

131,8

131,6

131,4

131,4

132,2

133,7

135,2

2014

2015

2016

2017

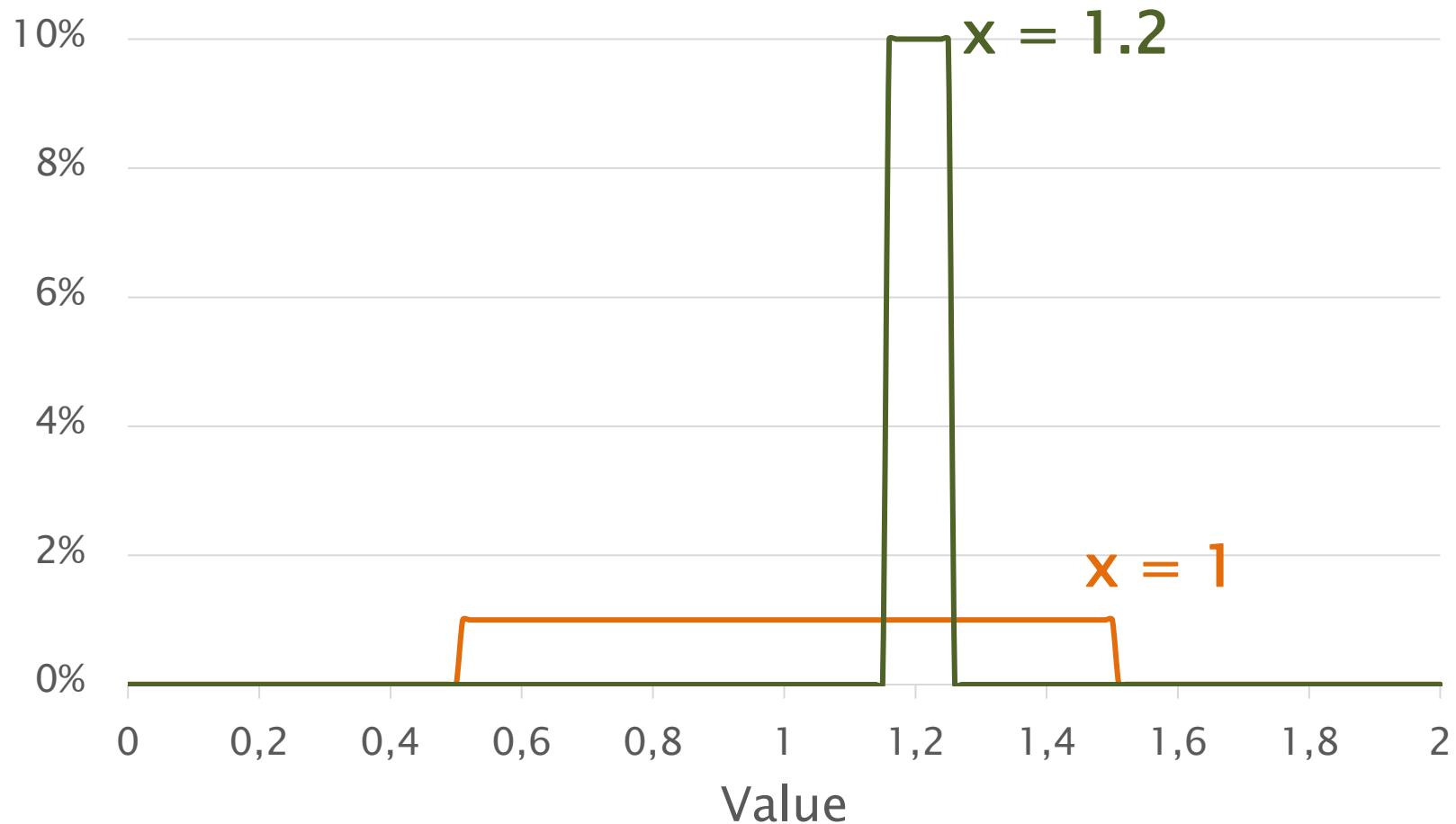
2018

2019 (*)

2020 (*)

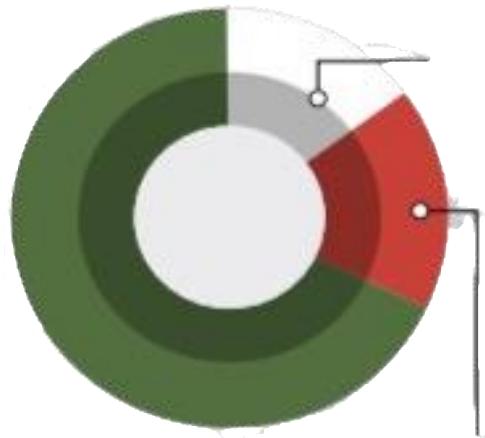
Precision and uncertainty

Probability

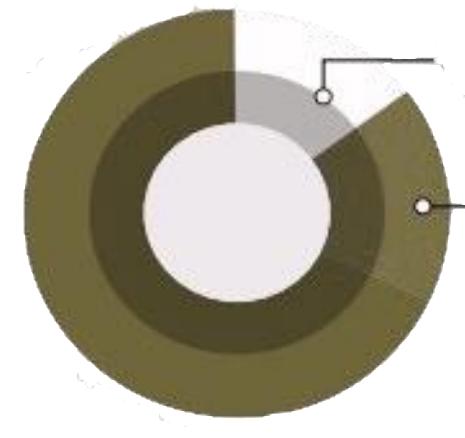


Accessibility

- The capability of data to be accessed, particularly by people who need supporting technology or special configuration because of some disability



Original



Color-blind simulation

References

- ISO/IEC 25010 – System and software quality models
- ISO/IEC 25012 – Data Quality model
- ISO/IEC 25024 – Measurement of data quality