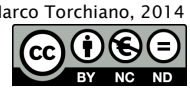
Data description and visualization

Empirical Methods in Software Engineering

http://softeng.polito.it/EMSE/

SoftEng
http://softeng.polito.it

Version 1.8



creative commons

Agenda

- Quantitative summaries
- Tables and Graphs
- Visual perception

Terms

- Diagram noun
  - a simplified drawing showing the appearance, structure, or workings of something; a schematic representation
  - ORIGIN early 17th cent.:
    - from Latin *diagramma, from Greek, from diagraphein 'mark out by lines,' from dia 'through' + graphein 'write.'*
- Graph [1] noun
  - a diagram showing the relation between variable quantities

Oxford American Dictionary

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

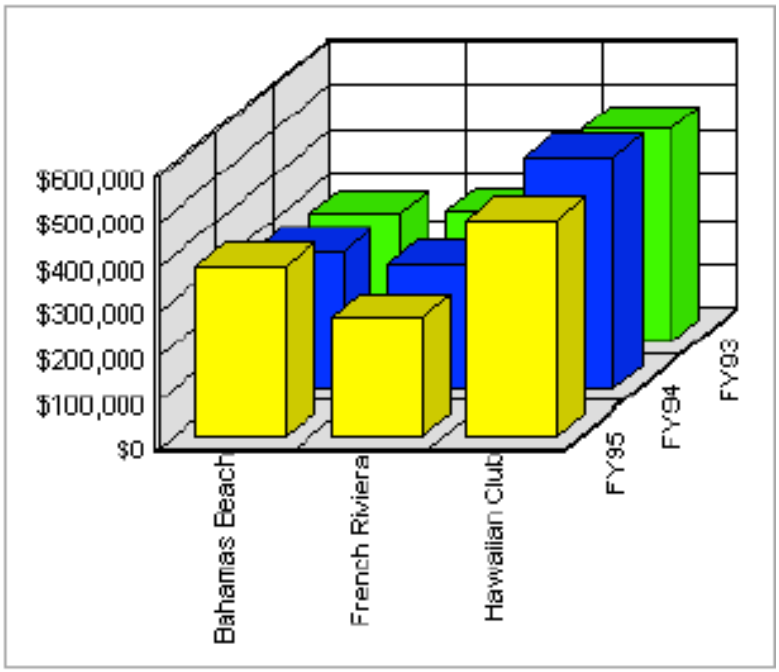
Understanding → Good Decisions

Information Visualization
Visual Patterns, Trends, Exceptions

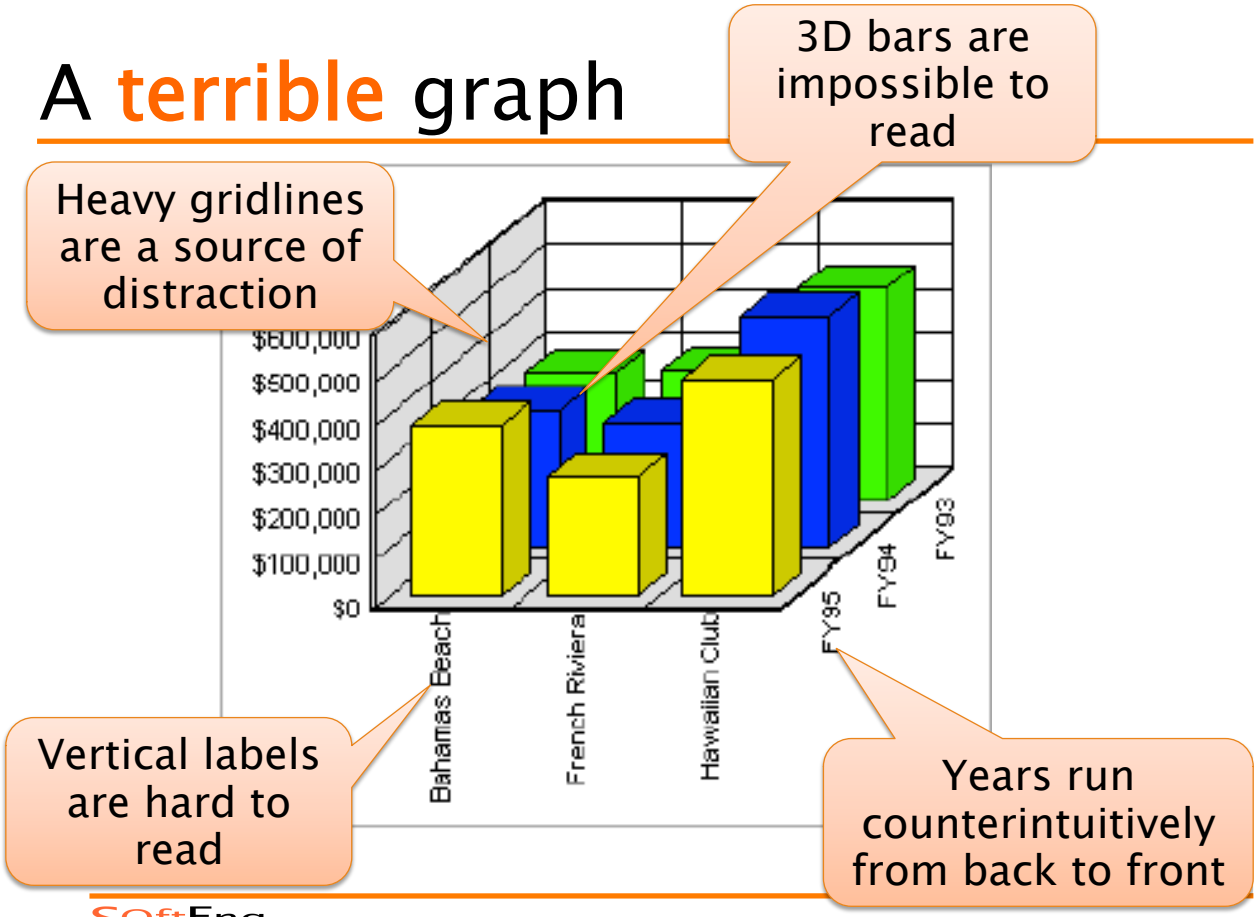
Quantitative Reasoning
Quantitative Relationships
Quantitative Comparisons

Visual Perception
Visual Properties
Visual Objects

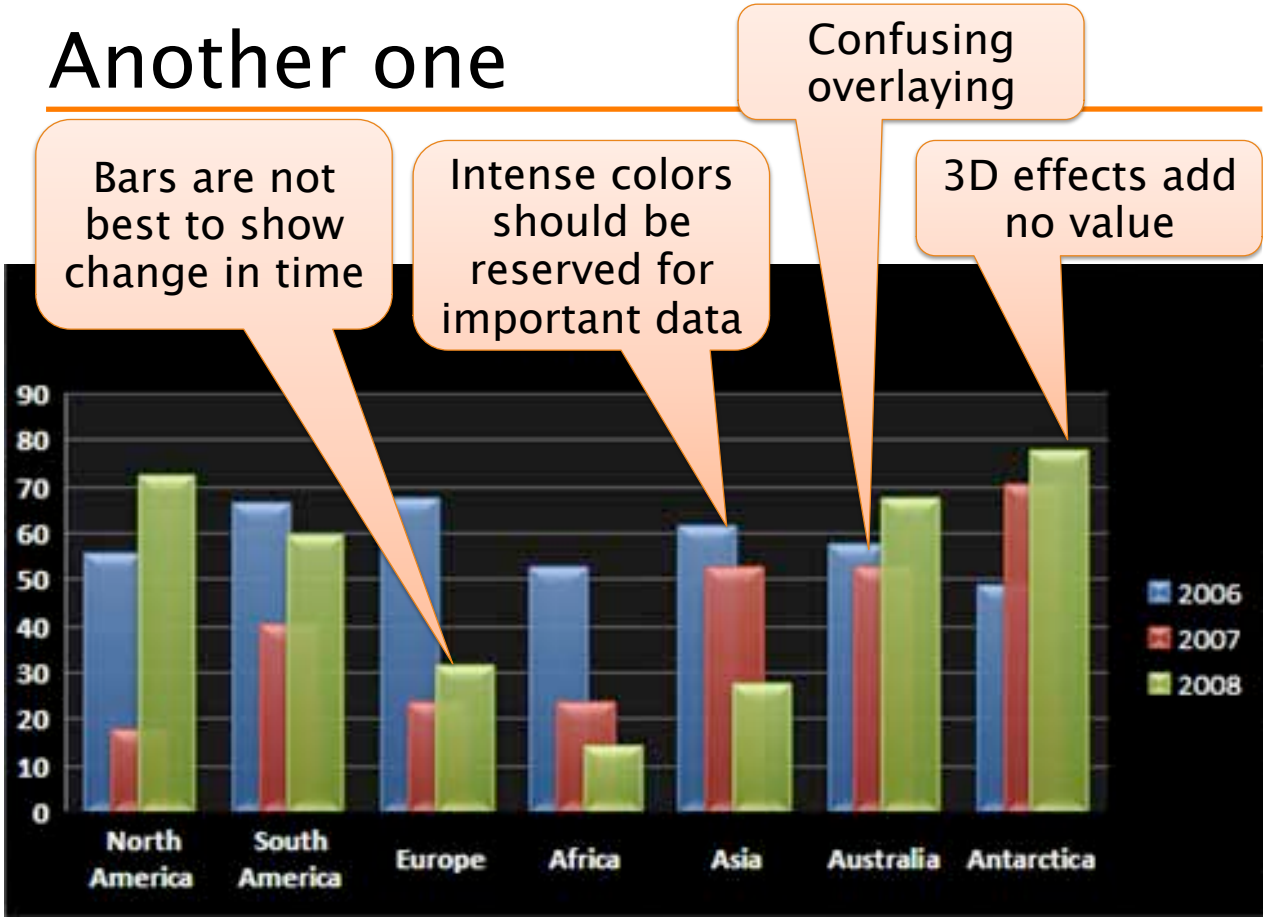
A graph



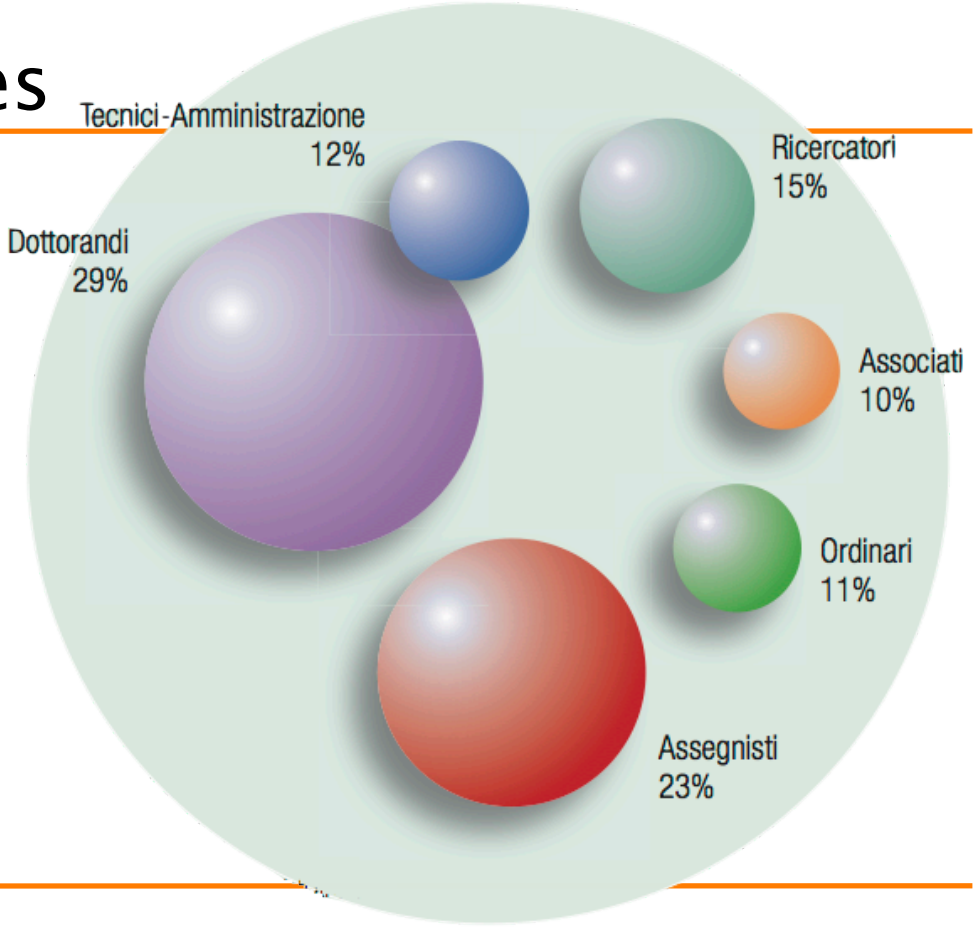
A terrible graph



Another one



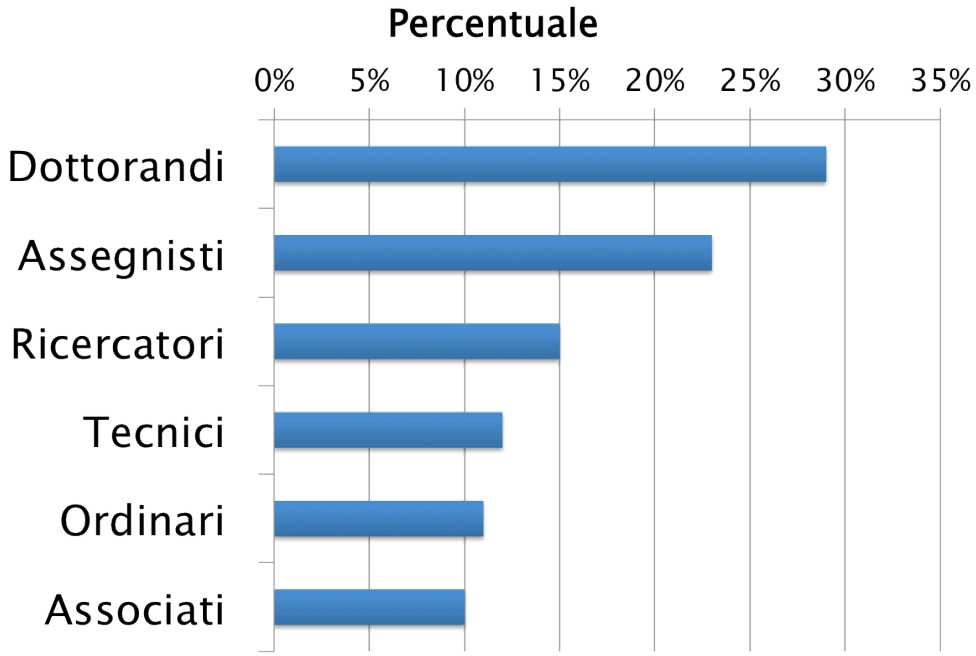
Bubbles



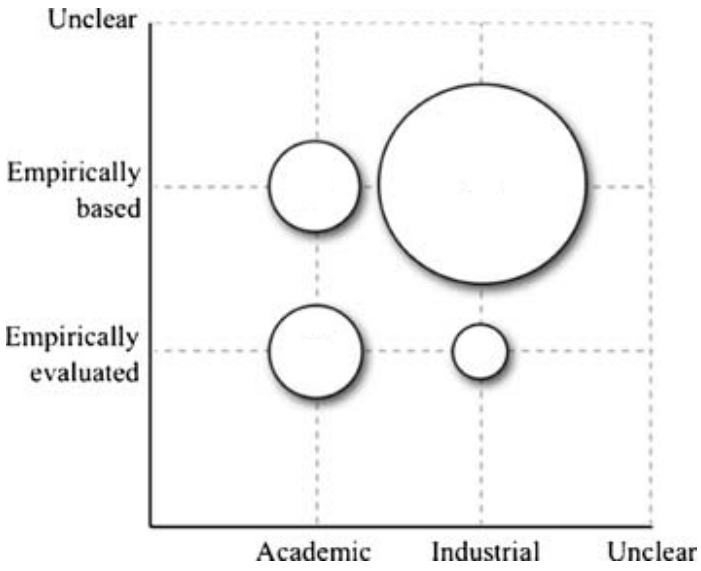
Analysis tasks

- How much larger is the proportion of PhD (Dottorandi) vs. Full professors (Ordinari)?
- Which is the fourth largest group of research related people?

Bar graph



Bubble plot



QUANTITATIVE SUMMARIES

Quantitative relationships

- Quantitative information conveys a message about relationships
  - Between quantitative and categorical
  - Among sets of quantitative

Categorical information

- Nominal
- Ordinal
  - Range
    - set of intervals on a ratio scale
  - Hierarchical
    - nested nominal categories

Categorical relationships

- Categories relate to each other through quantitative values associated with them
  - Ranking
  - Proportion
  - Correlation

Summarizing data

- Central tendency (average)
  - Mean (arithmetic)
  - Median
  - Mode
  - Midrange
- Dispersion
  - Range
  - Standard Deviation

Summarizing data

- Correlation
  - Correlation coefficients
    - Does correlation exists?
    - How strong/weak is it?
    - Is it positive or negative?
  - Scatter plots
- Proportion
  - Fraction
  - Rate
  - Percentage

Money

- Raw values
- Adjusted for inflation

QUANTITATIVE COMMUNICATION

Quantitative message

- Quantitative values
  - Express measures

- Categories
  - Identify what the values refer to
  - Entities groups

Quantitative message

Product	Pre-release	Post-release
Word processor	13,248	1,724
Spreadsheet	9,487	956
Presentation	7,432	1,045
Total	30,167	3,725

Quantitative message

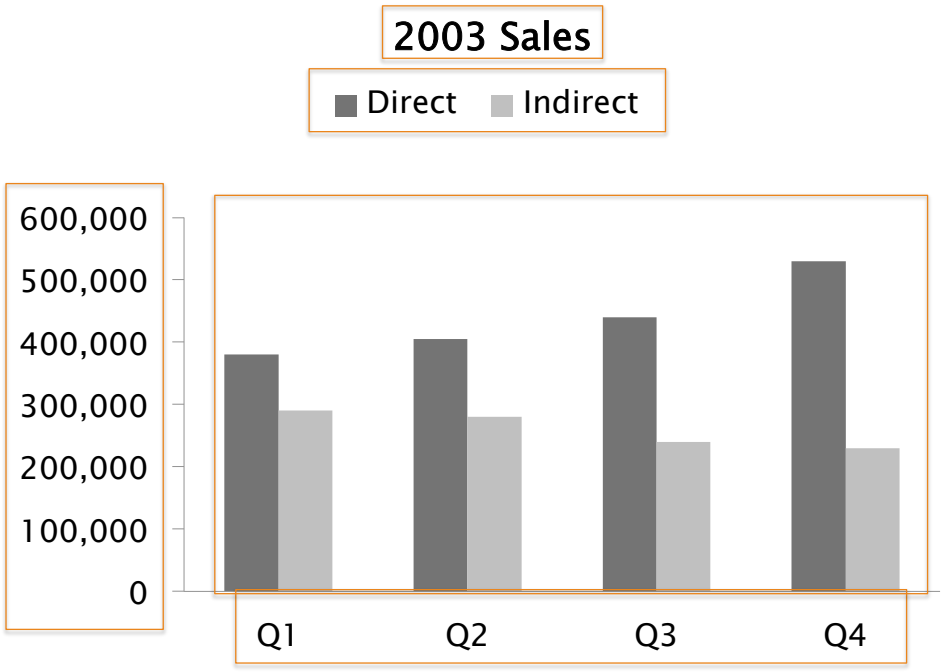
Product	Pre-release	Post-release
Word processor	13,248	1,724
Spreadsheet	9,487	956
Presentation	7,432	1,045
Total	30,167	3,725

Categorical subdivisions

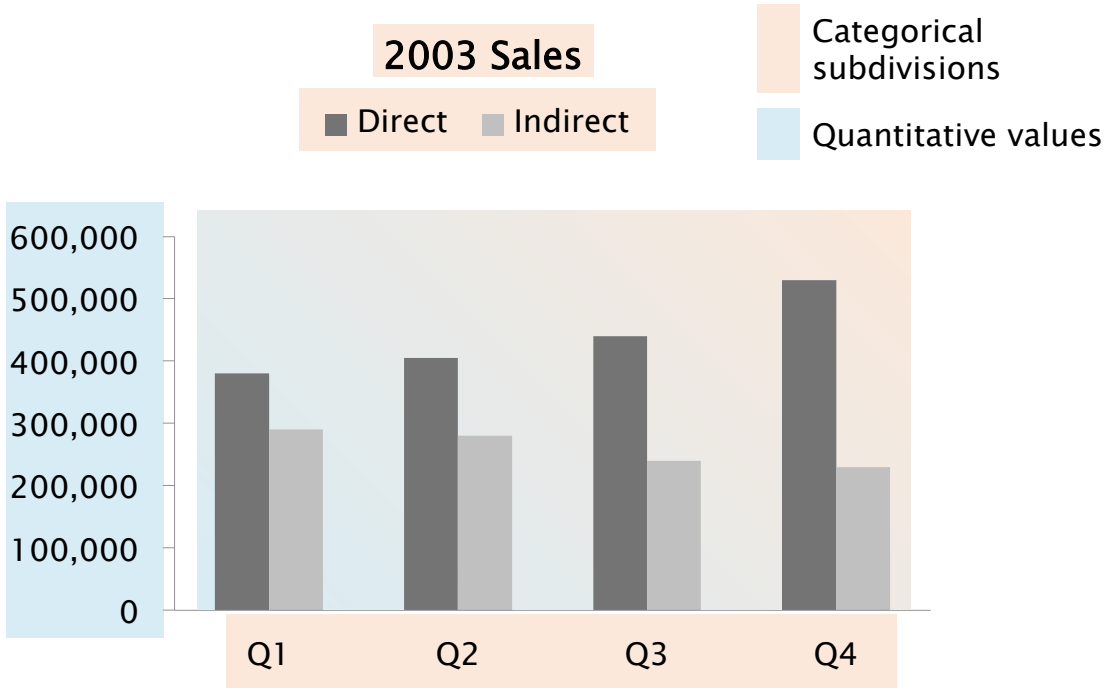
Category name

Quantitative values

Quantitative message



Quantitative message



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	Use graphs to
Look up individual values	Focus on the shape of values
Compare individual values	Reveal relationships among multiple values
Precise values are required	
There is more than one unit of measure	

GRAPHICAL INTEGRITY

Principles of integrity

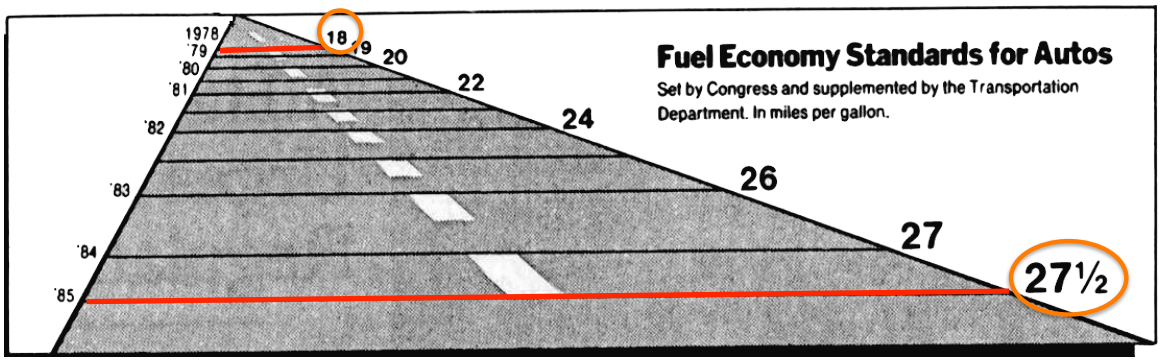
- Proportionality
  - Representation as physical quantities should be proportional to the represented numbers
- Utility
  - Graphical element should convey useful information
- Clarity
  - Labeling should counter graphical distortion and ambiguity

Lie Factor

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

- Overstating
  - Log(LF) $>$ 0
- Understating
  - Log(LF) $<$ 0

Lie Factor



$$\frac{18.7}{2.2} = 8.5 \text{ on graphic} \qquad \frac{27.5}{18} = 1.52 \text{ in data}$$

$$LF = 8.5 \,/\, 1.52 = 5.59$$

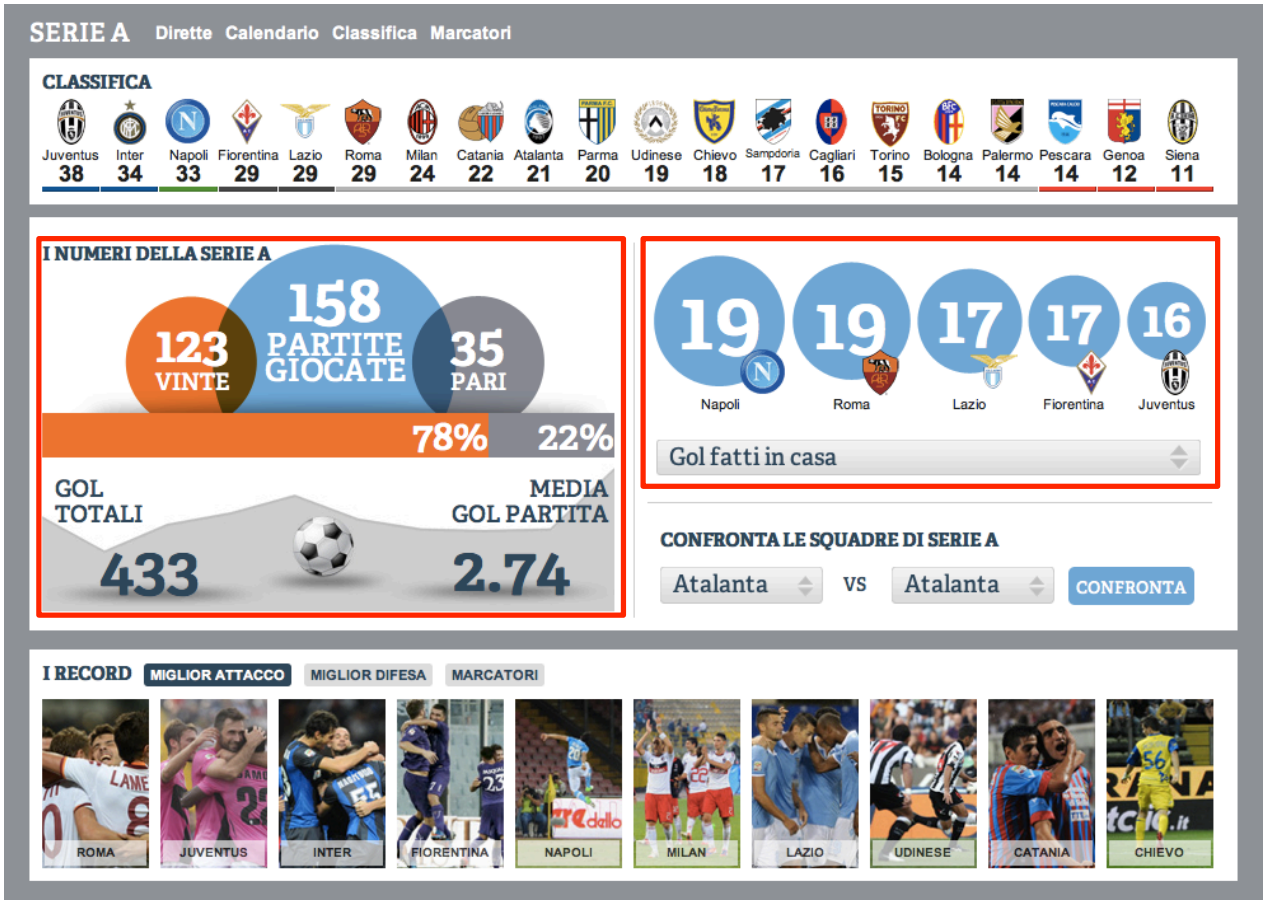
Data-ink

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

- Proportion of a graphic's ink devoted to the non-redundant display of data information
- 1 – proportion of a graphic that can be erased without loss of information

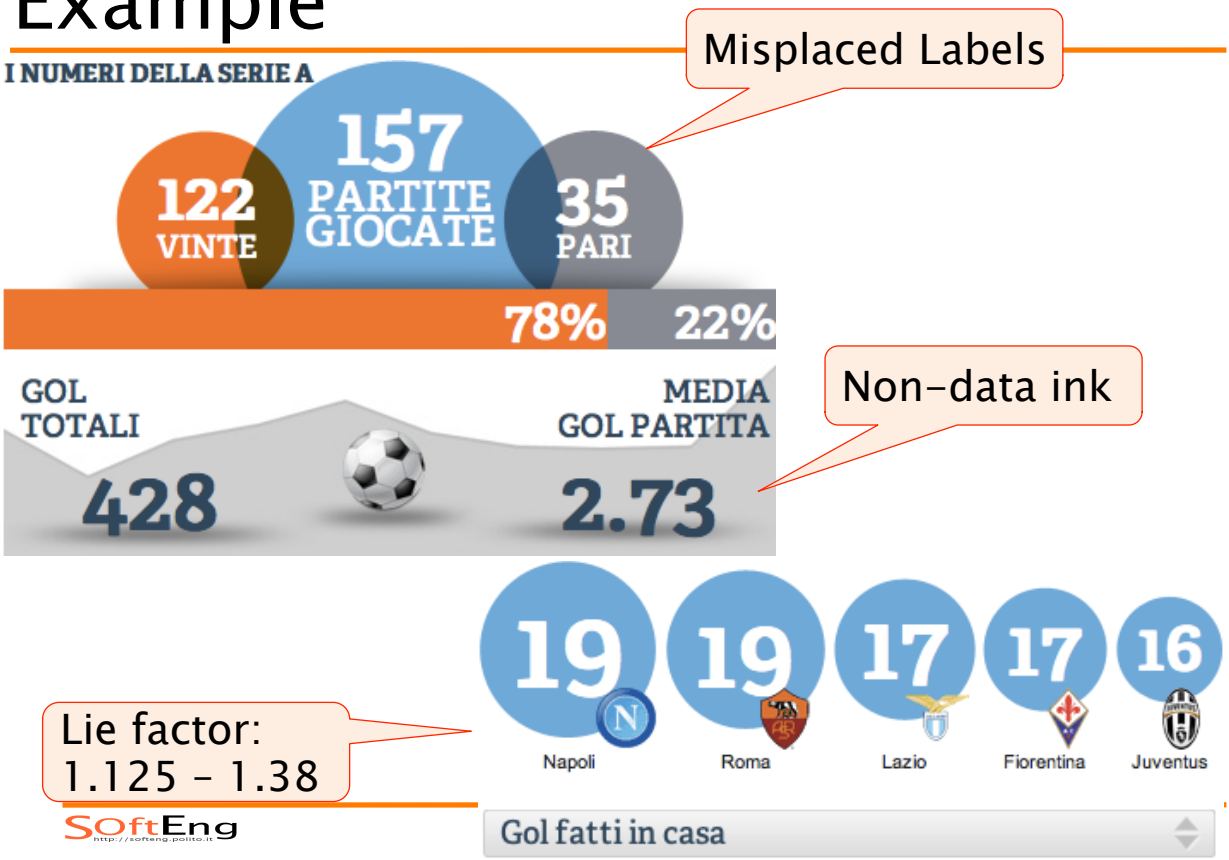
Principles of design

- Maximize data-ink ratio
  - Erase non-data-ink
  - Erase redundant data-ink
- Within reason



 URL: http://sport.repubblica.it/classifica/A , visited on December 10, 2012

Example



TABLES

Data encoding

- Categories and values are encoded in textual form

Relationships

- Categorical-to-Quantitative (look up)
  - Single set of C levels and Single set of Q values
  - Intersection of multiple C and Single set of Q values
- Quantitative-to-Quantitative (comparison)
  - Single set of Q values associated with different C levels
  - Distinct set of Q values associated with the same C level

Look up: 1C-1Q

Product	Defects
Word processor	13,248
Spreadsheet	9,487
Presentation	7,432
Total	30,167

Look up: 1C-1Q

Product	Defects
Word processor	13,248
Spreadsheet	9,487
Presentation	7,432
Total	30,167

Look up: nC-1Q

Product	Pre-release	Post-release
Word processor	13,248	1,724
Spreadsheet	9,487	956
Presentation	7,432	1,045
Total	30,167	3,725

Look up: nC-1Q

Product	Pre-release	Post-release
Word processor	13,248	1,724
Spreadsheet	9,487	956
Presentation	7,432	1,045
Total	30,167	3,725

Comparison: 1Q

Product	Pre-release	Post-release
Word processor	13,248	1,724
Spreadsheet	9,487	956
Presentation	7,432	1,045
Total	30,167	3,725

Comparison: 1Q

Product	Pre-release	Post-release
Word processor	13,248	1,724
Spreadsheet	9,487	956
Presentation	7,432	1,045
Total	30,167	3,725

Comparison: nQ

Product	Defects	Test Effort
Word processor	13,248	300
Spreadsheet	9,487	600
Presentation	7,432	500
Total	30,167	1,400

Comparison: nQ

Product	Defects	Test Effort
Word processor	13,248	300
Spreadsheet	9,487	600
Presentation	7,432	500
Total	30,167	1,400

Design variation

- Unidirectional
  - Categorical levels are laid out in one direction only
    - Across columns or rows
  - Possibly in hierarchical arrangement
- Bidirectional
  - Categorical levels are laid out in both directions

Suitable designs

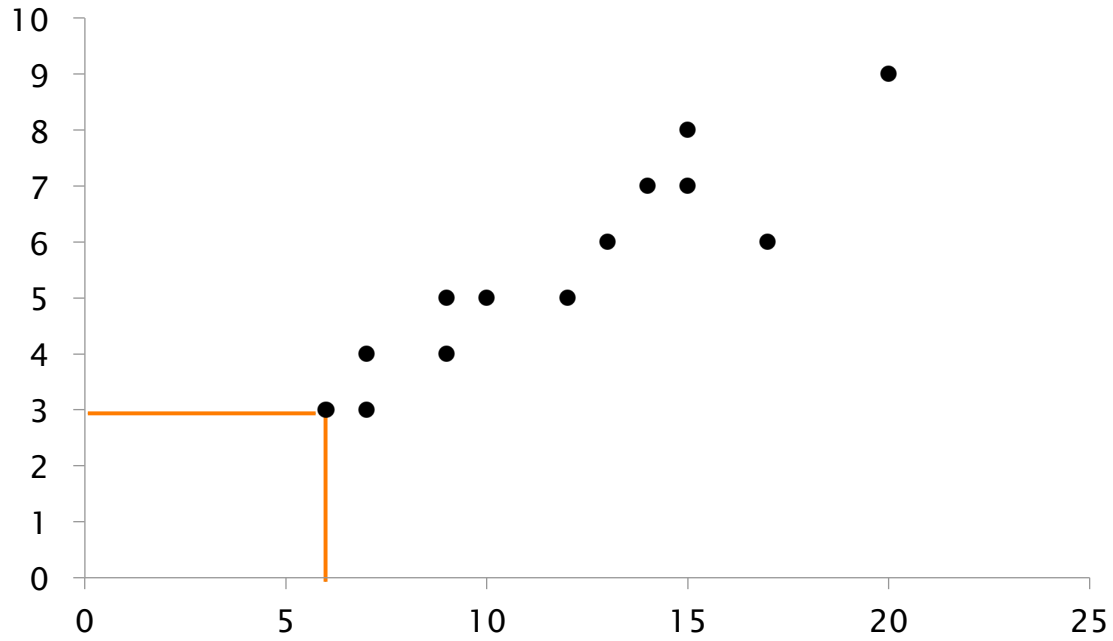
		Unidirectional	Bidirectional
C-to-Q Look up	1C-1Q	✔	*NA*
	nC-1Q	~	✔
Q-to-Q Comparison	1Q	✔	✔
	nQ	✔	~

GRAPHS

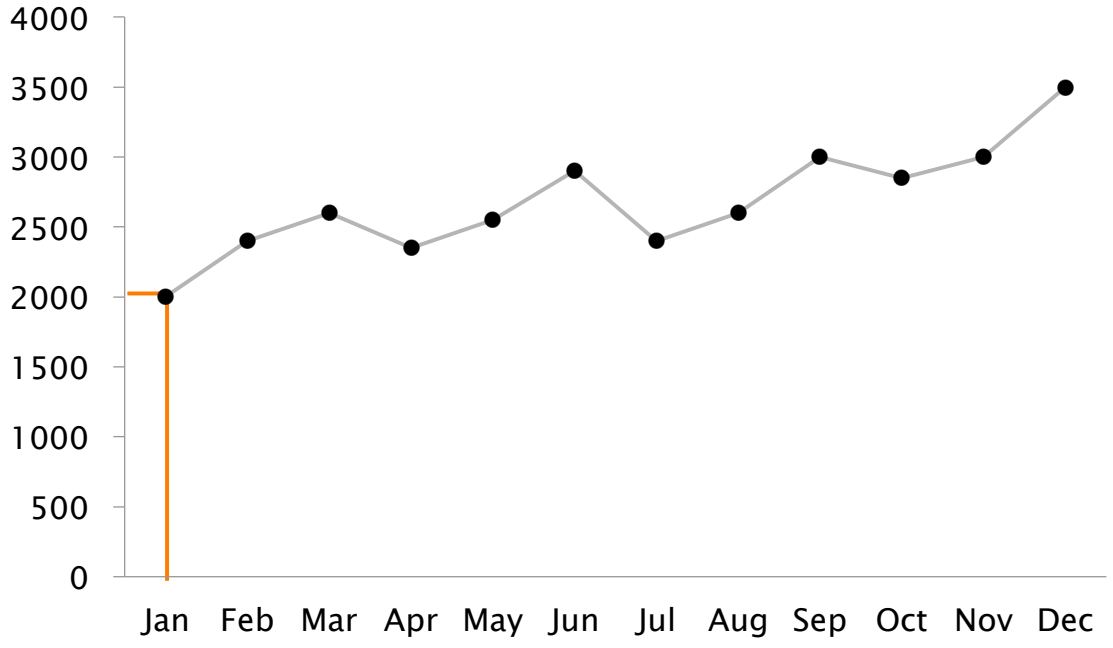
Quantitative encoding

- Encoding of quantitative values
  - Points -> relative position
  - Lines -> relative position, slope, length
  - Bars -> length
  - Shapes -> 2D areas

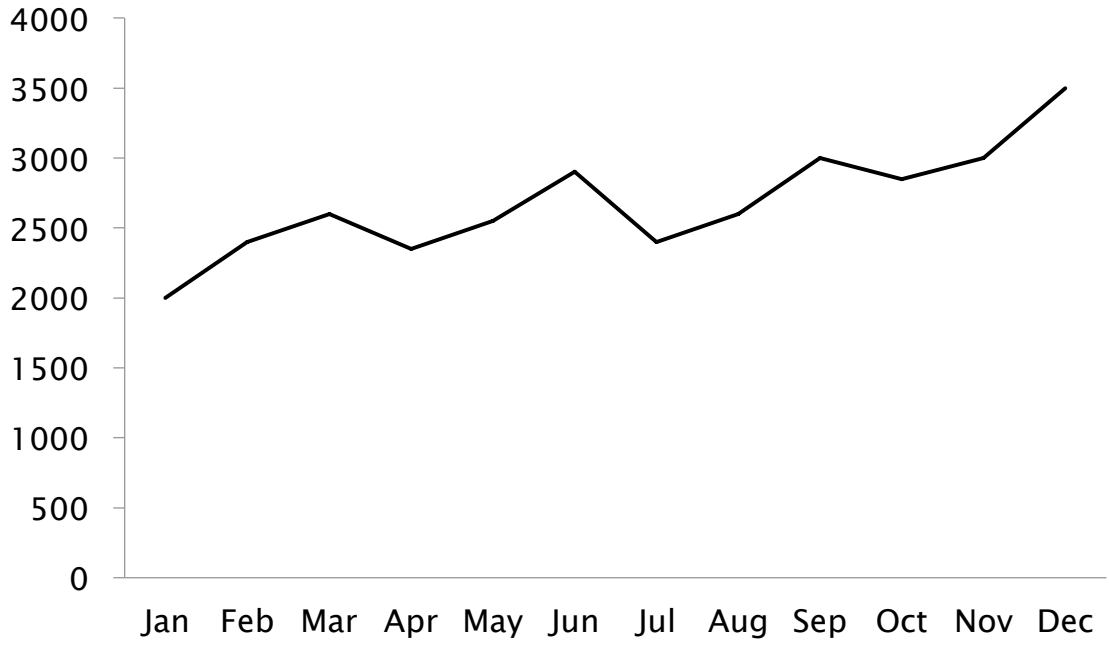
Points



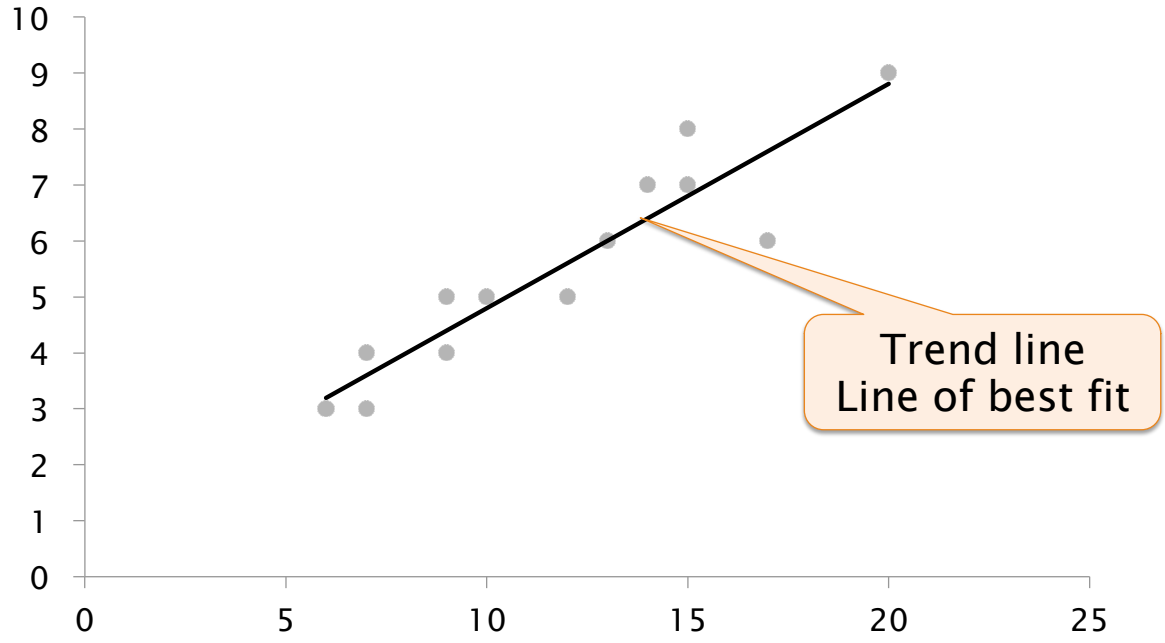
Points and Lines



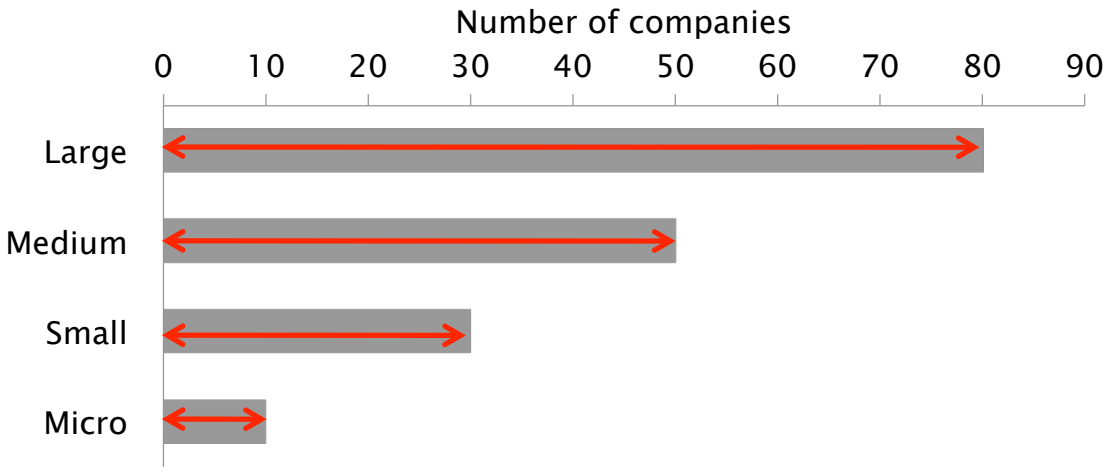
Lines


4000
3500
3000
2500
2000
1500
1000
500
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec

Lines


10
9
8
7
6
5
4
3
2
1
0
0
5
10
15
20
25
Trend line
Line of best fit

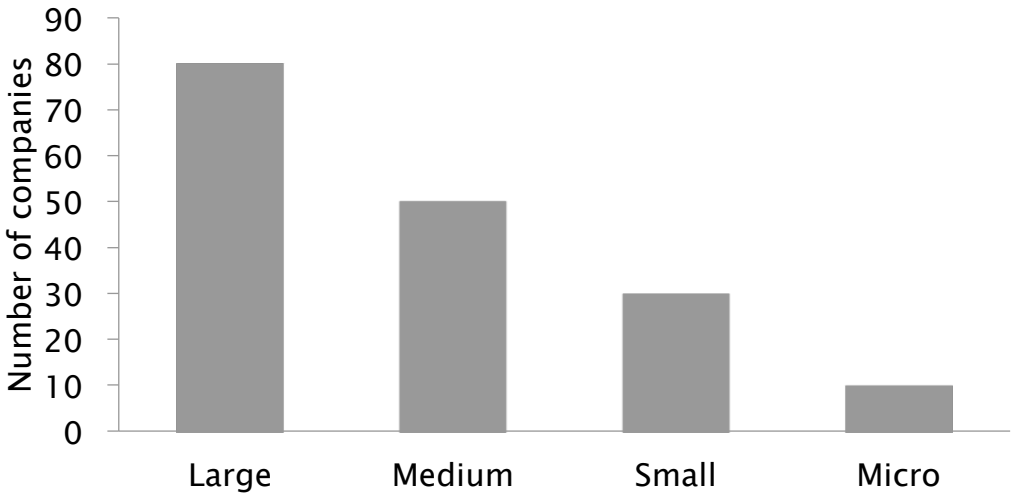
Bars



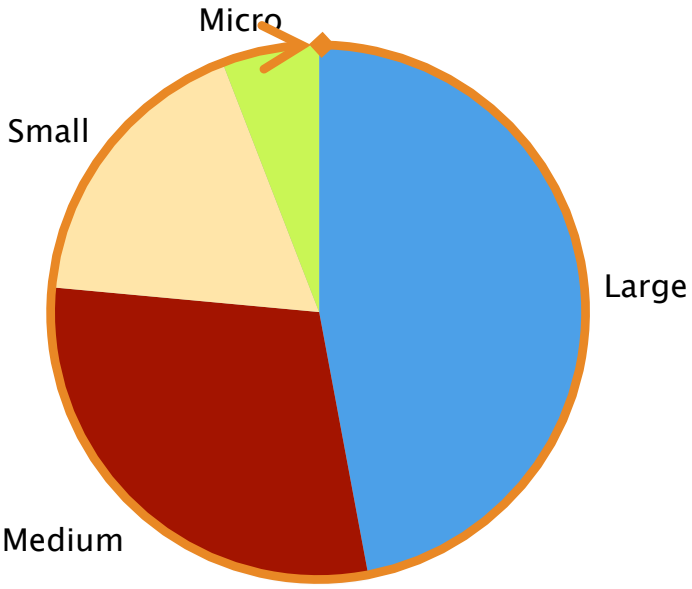
Bars – Warning

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

Vertical Bars (Columns)



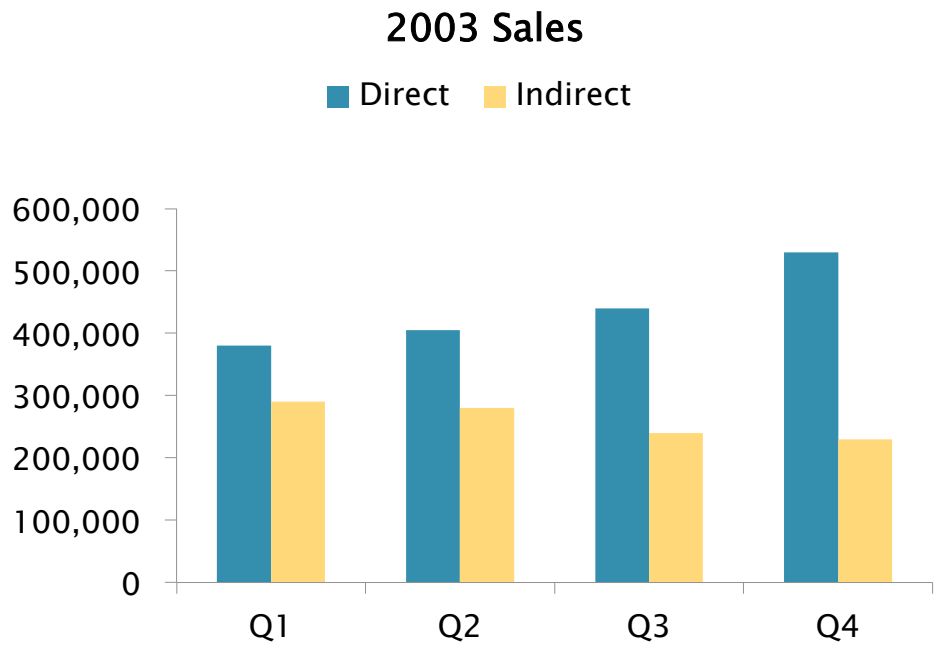
Area of 2D Shapes



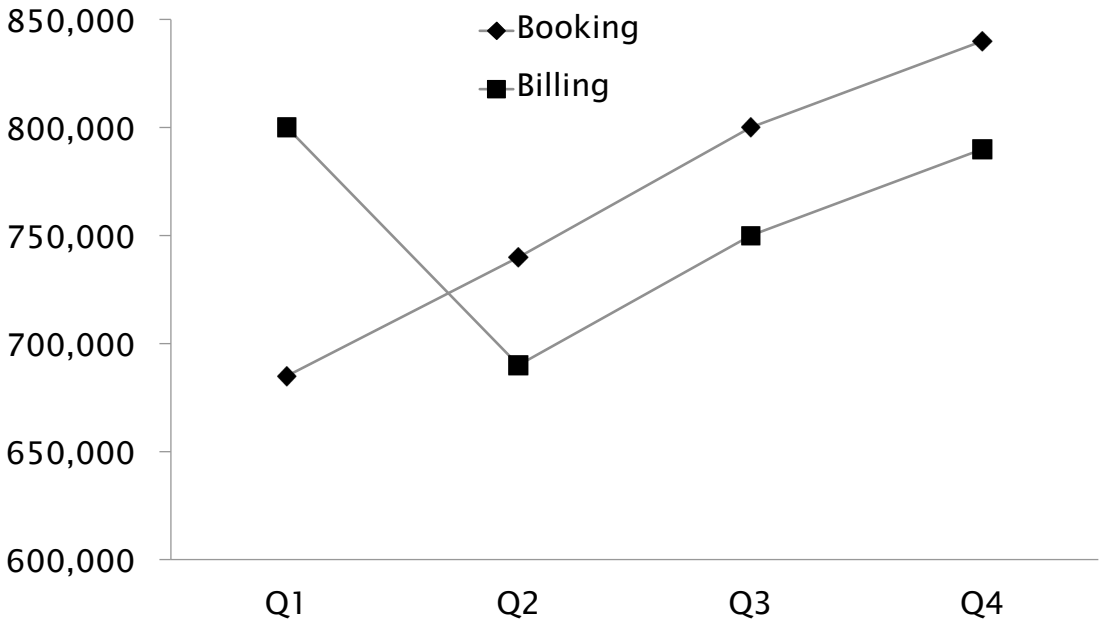
Data encoding

- Encoding of categorical levels
  - Position (along an axis)
  - Color
  - Shape
  - Fill pattern
  - Line style

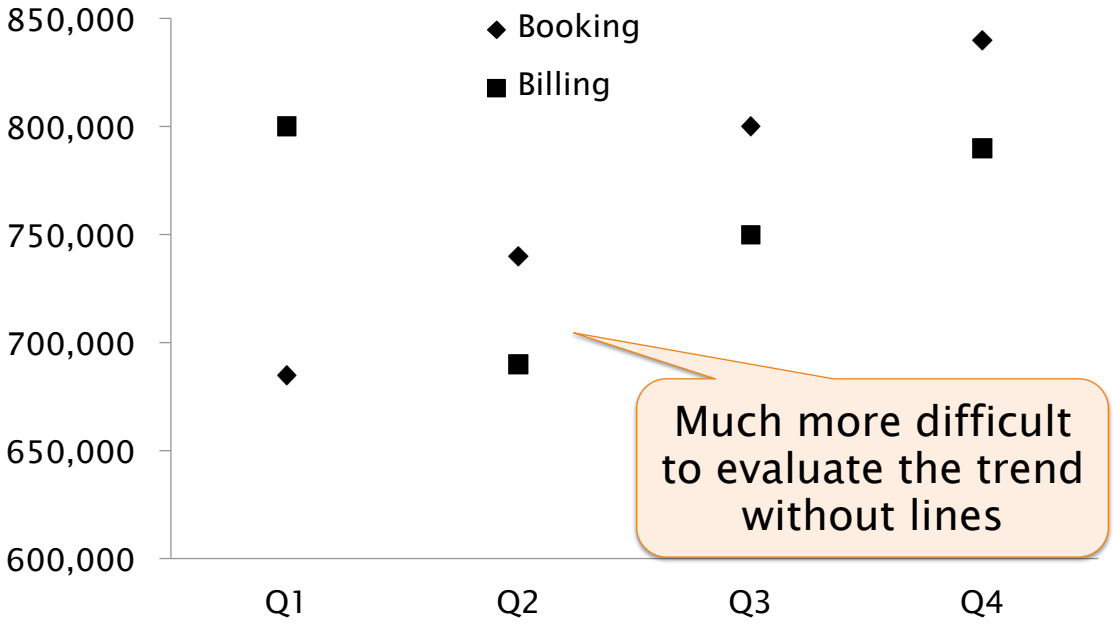
Quantitative message



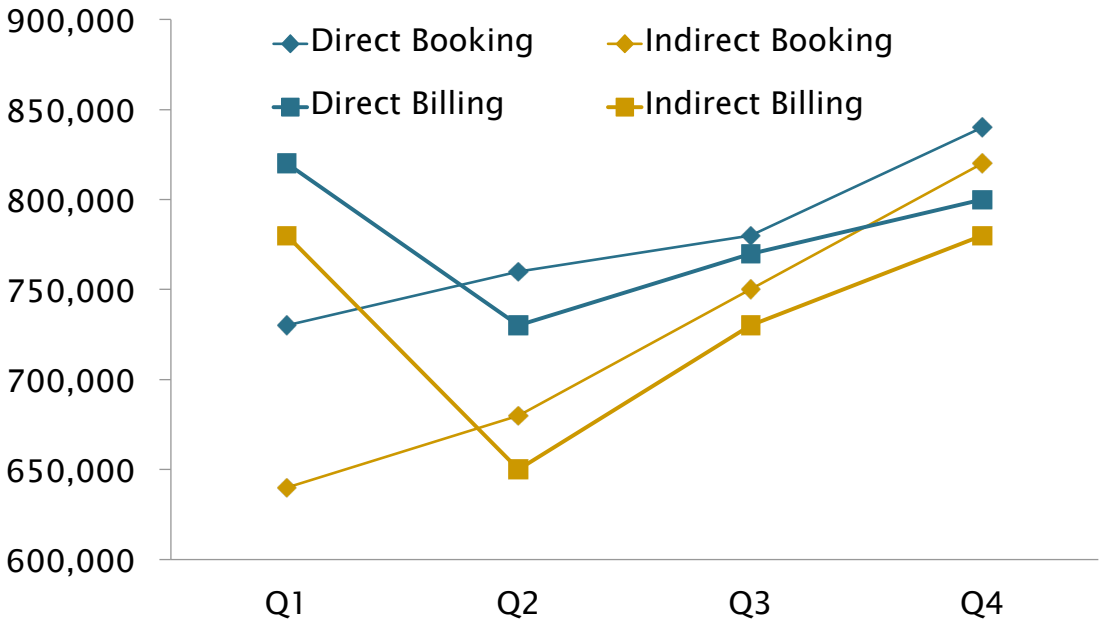
Point shape



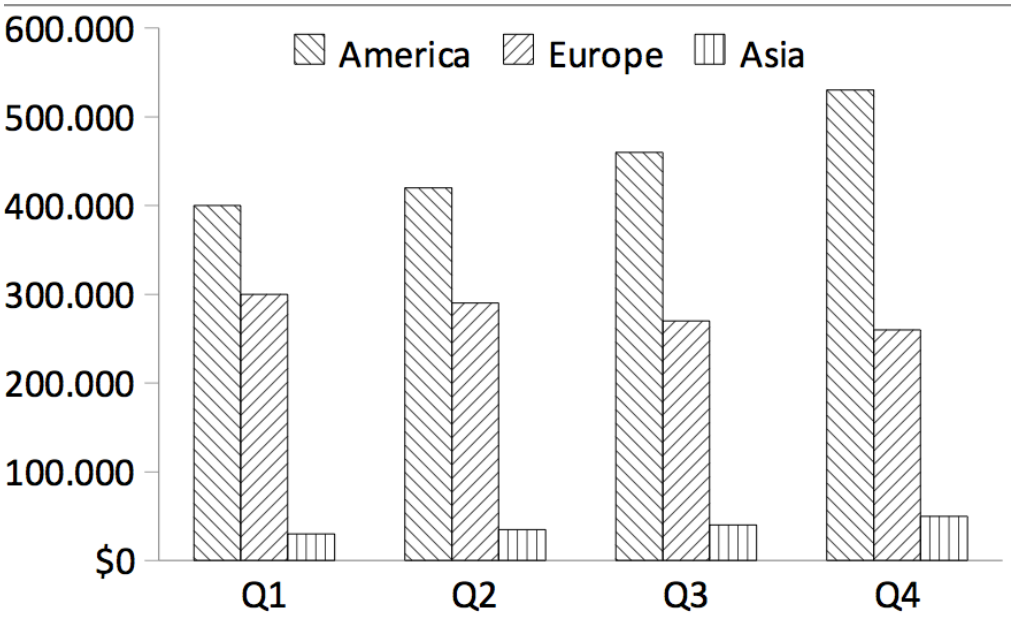
Point shape



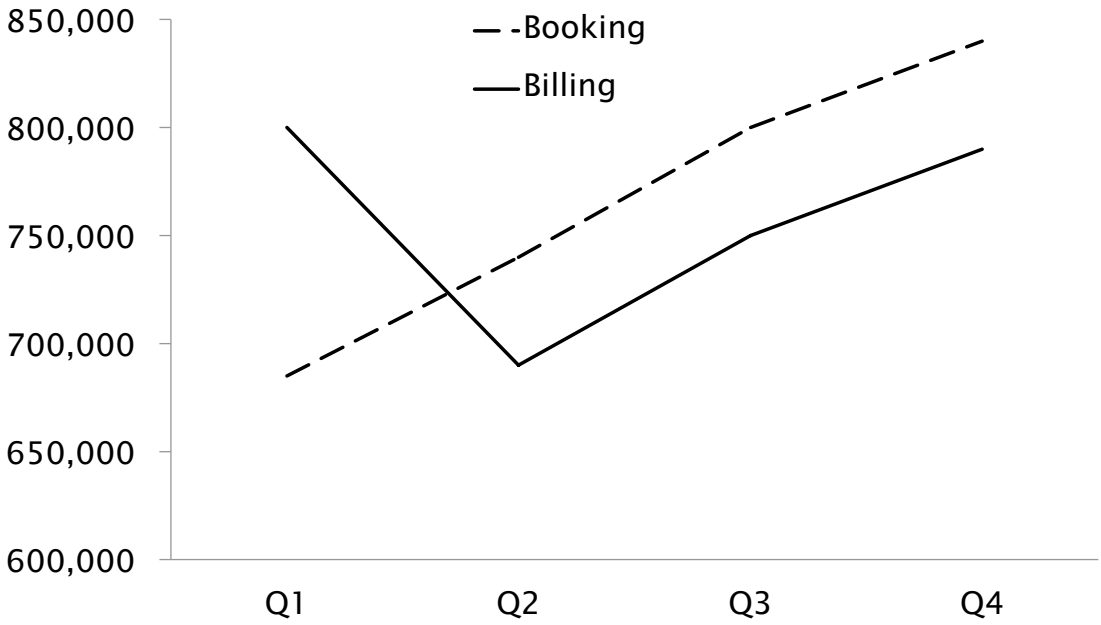
Point shape + Color



Fill Pattern



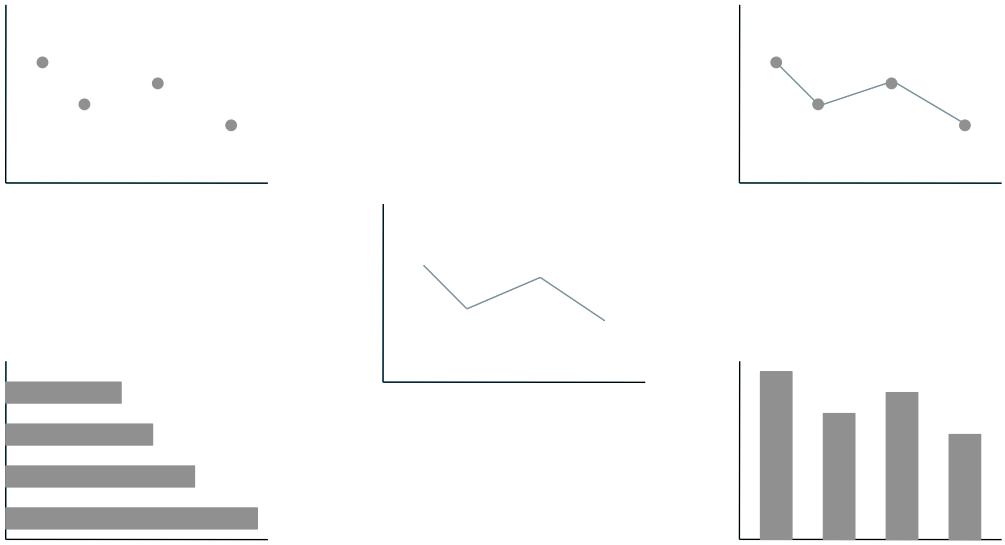
Line style



Relationships in graphs

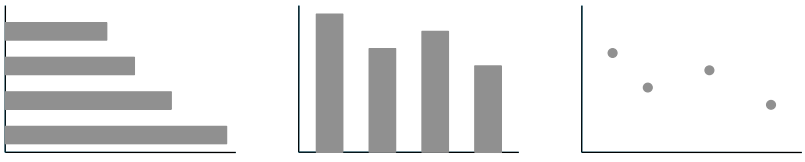
- Nominal comparison
- Time series
- Ranking
- Part-to-whole
- Deviation
- Distribution
- Correlation

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

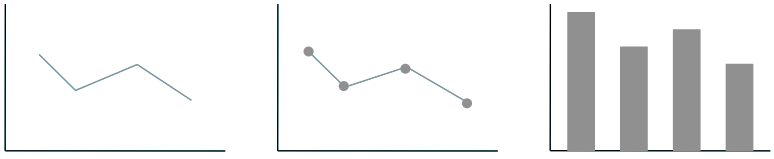


Time series

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

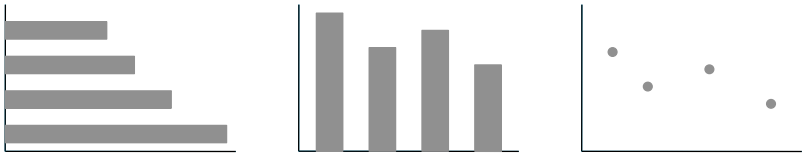
Time series

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



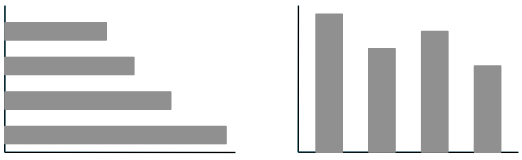
Ranking

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

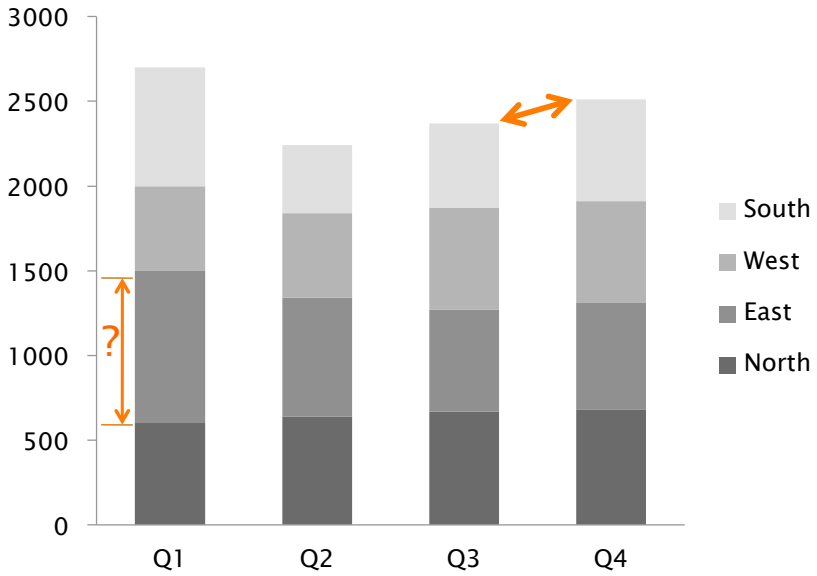


Part-to-whole

- Best unit: percentage
- Stacked bar graph
  - Difficult to read individual values

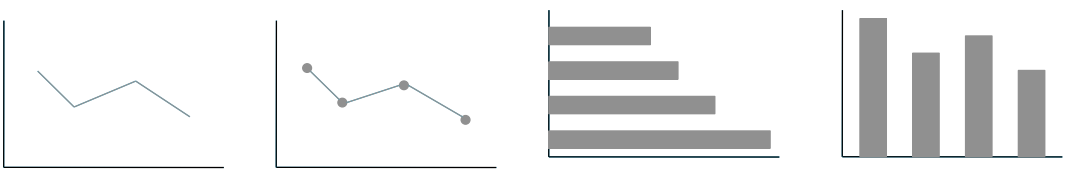


Stacked bar graph

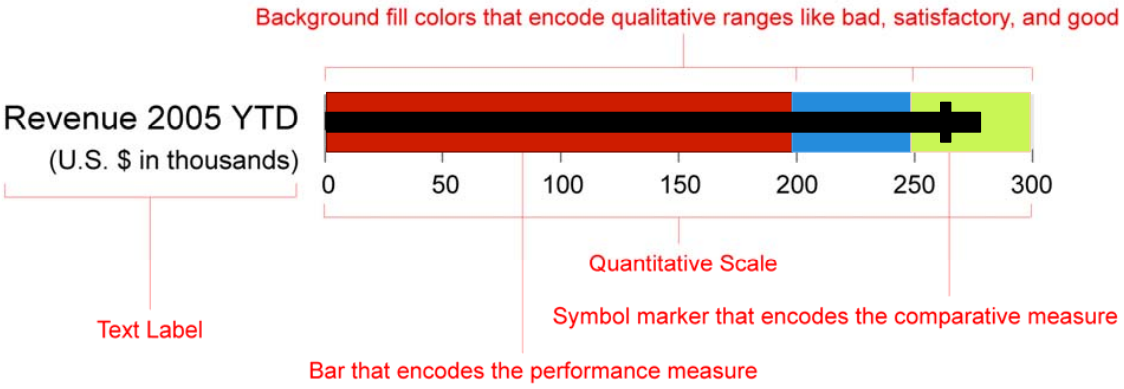


Deviation

- To what degree one or more sets of values differ in relation to primary values.
  - Often linked to time series



Bullet graph



Distribution

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

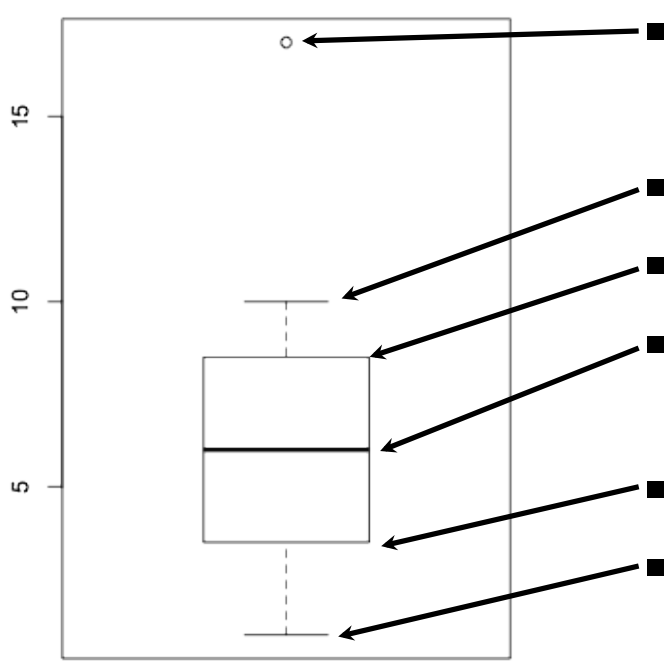
Single distribution

- Histogram
  - Vertical bar graph
  - Frequency for subdivision
    - Quantitative ranges
    - Categories
- Emphasis on number of occurrences in each subdivision

Single distribution

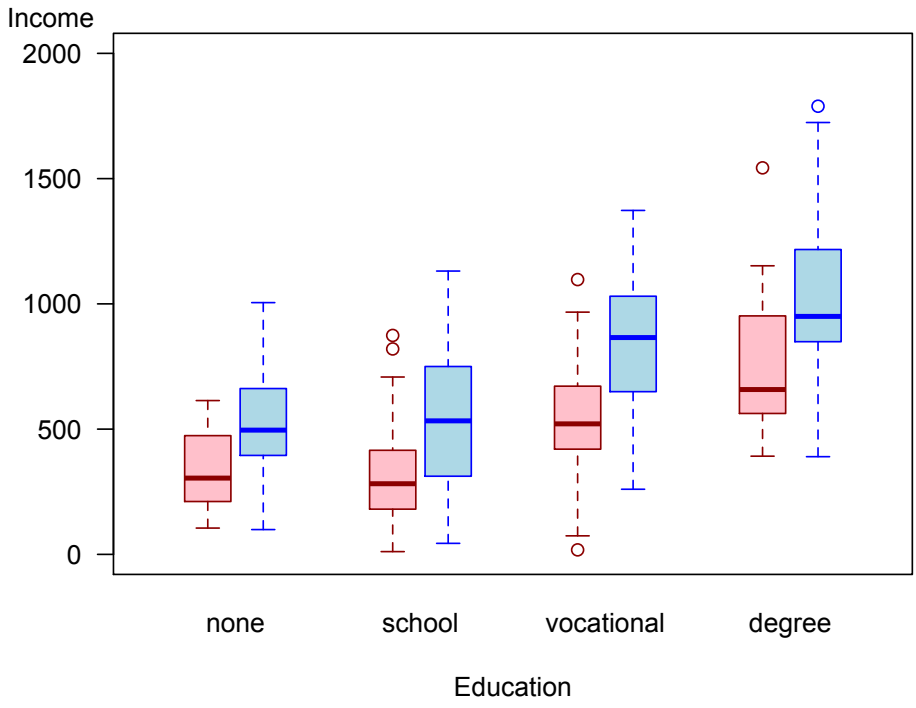
- Frequency polygon
  - Line graphs
  - Frequency density function
- Emphasis on the shape of the distribution

Box plot

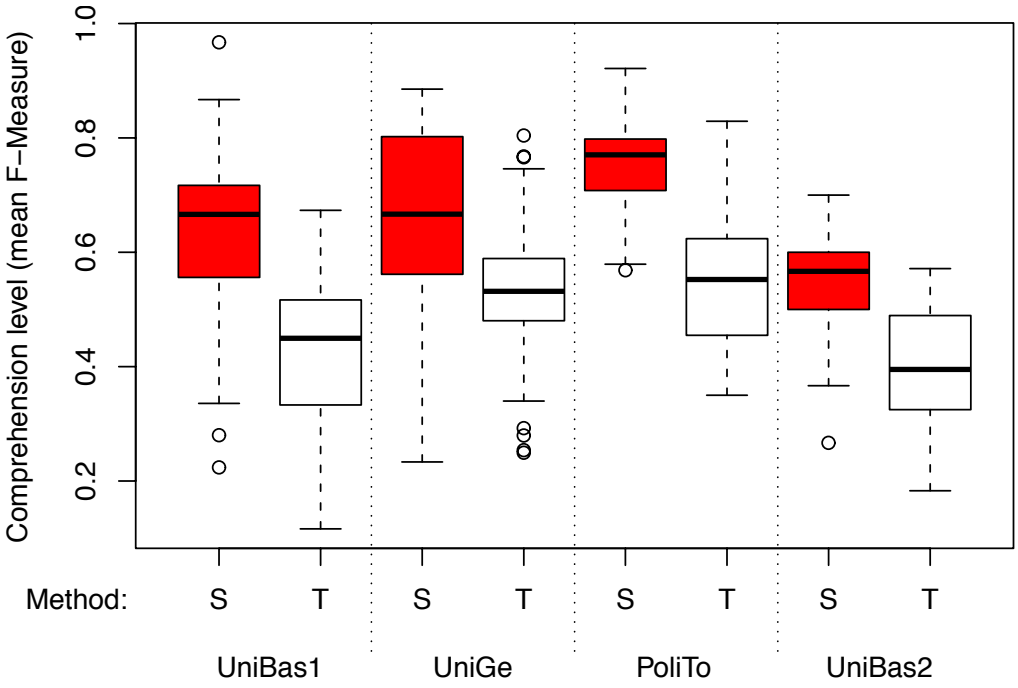


- Oulier
- Max value
- 75th percentile
- Median
  - 50th percentile
- 25th percentile
- Min value

Box plot



Boxplot

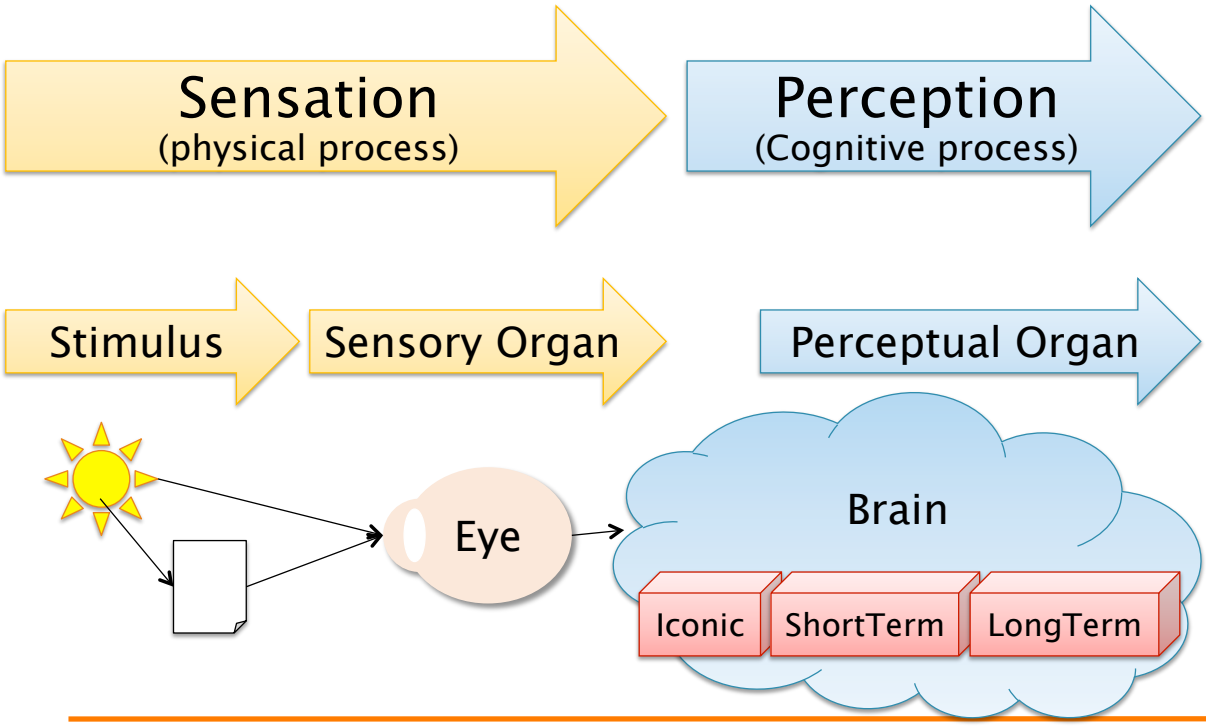


Correlation

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

VISUAL PERCEPTION

Visual perception



Memories

- 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

Attention blindness

http://www2.psych.ubc.ca/~rensink/flicker/download/index.html

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

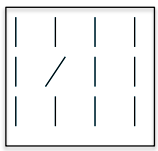
Pre-Attentive Attributes

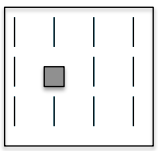
5 7 8 4 9 8 3 1 1 0 6 8 8 2 1 1 5 2 6 6 5
9 5 1 8 4 6 8 4 9 3 0 4 5 3 4 9 2 5 8 5 8
5 0 5 4 6 2 6 5 7 3 7 8 6 5 3 7 2 6 3 1 5
5 8 6 6 8 3 7 6 5 0 9 6 3 4 6 1 9 5 6 6 4
1 6 7 3 9 9 2 8 3 4 0 3 5 1 6 3 5 3 9 3 4
8 6 9 7 5 4 2 4 7 4 9 5 8 5 3 0 7 6 0 6 7
0 3 1 5 3 2 3 5 6 7 2 8 9 8 5 3 7 8 8 2 4
5 5 3 4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

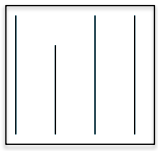
Pre-Attentive attributes

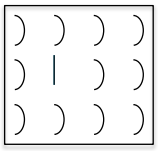
Category	Attribute
Form	Orientation Line length Line width Size Shape Curvature Added marks Enclosure
Color	Hue Intensity
Spatial position	2-D position
Motion	Flicker Direction

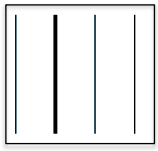
Attributes of form

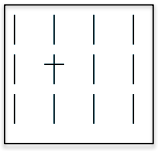
Orientation 

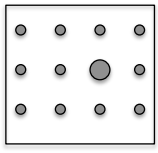
Shape 

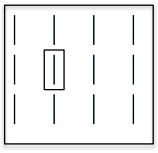
Line Length 

Curvature 

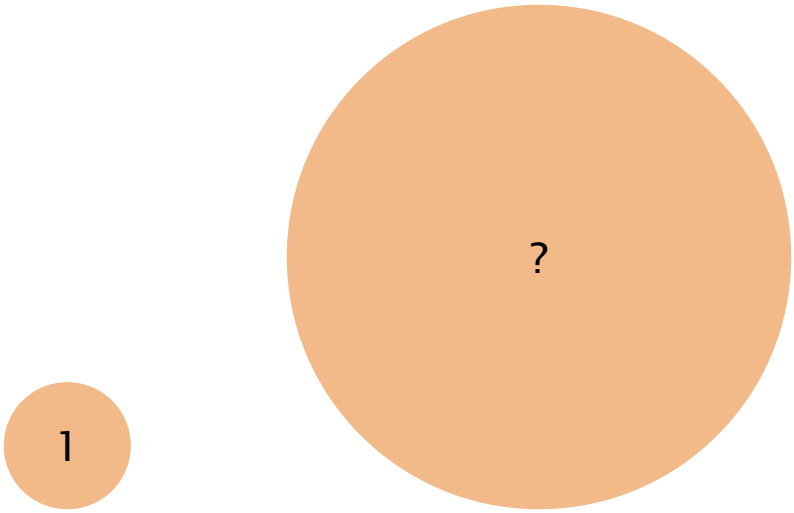
Line Width 

Added mark 

Size 

Enclosure 

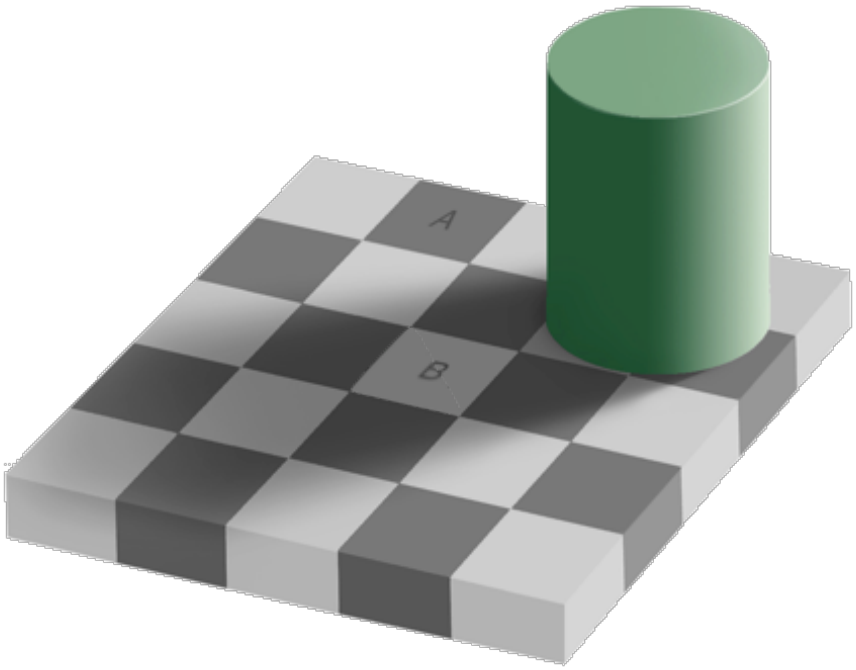
Size



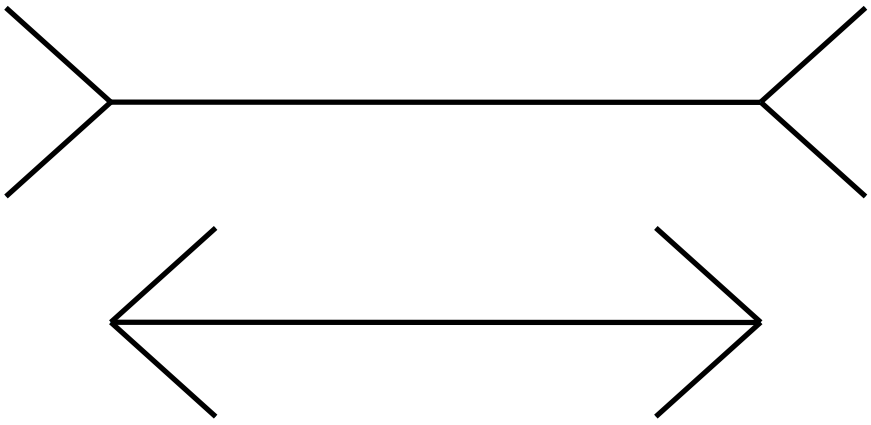
Pre-Attentive attributes

Category	Attribute	Quantitative
Form	Orientation	No
	Line length	Yes
	Line width	Limited
	Size	Limited
	Shape	No
	Curvature	No
	Added marks	No
	Enclosure	No
Color	Hue	No
	Intensity	Limited
Spatial position	2-D position	Yes
Motion	Flicker	No
	Direction	No

Effect of Context



Effect of context

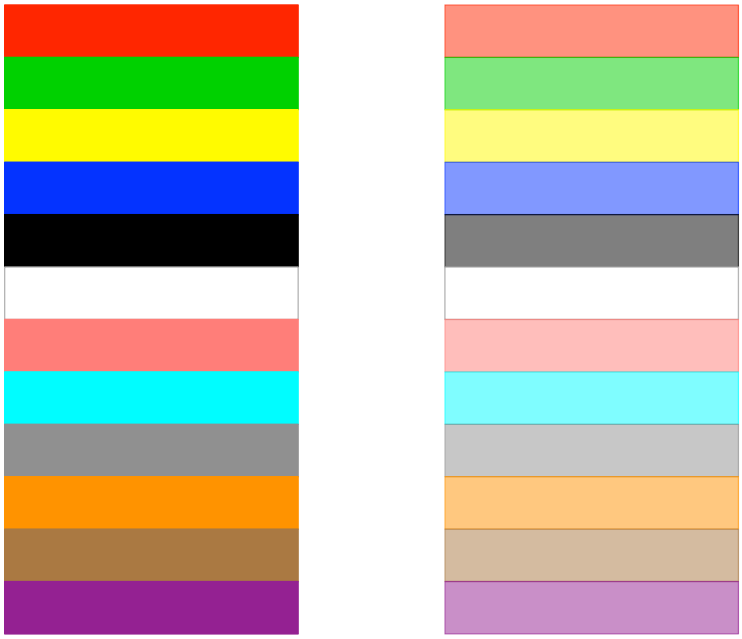


Distinct perceptions

- 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 attribute not more than four distinctions are discernible

Colors

- Red
- Green
- Yellow
- Blue
- Black
- White
- Pink
- Cyan
- Gray
- Orange
- Brown
- Purple



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

Gestalt principles

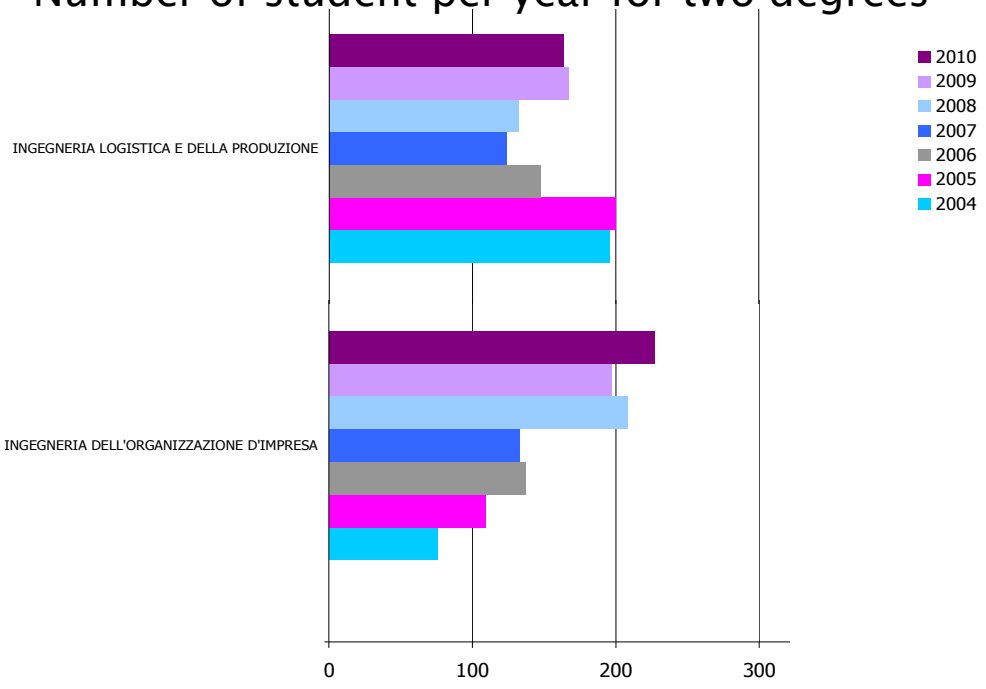
- Visual attributes that lead us to group the objects
  - Proximity
  - Similarity
  - Enclosure
  - Closure
  - Continuity
  - Connection

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)
  - Importance
  - Ordering
  - Time

Example

Number of student per year for two degrees



Prioritization

Attribute	Tables	Graphs
Line width	Boldface text	Thicker lines
Size	Bigger tables Larger fonts	Bigger graphs Wider bars Bigger symbols
Color intensity	Darker or brighter colors	
2-D position	Positioned at the top Positioned at the left Positioned in the center	

Analytical techniques

- Optimal quantitative scales
- Reference lines and regions
- Trellies
- Multiple concurrent views and brushing
- Focus and context together
- Details on demand
- Overplotting reduction

TABLE DESIGN

Table design

- Data components
  - Categorical level
  - Quantitative value
  - Complementary text
- Support components
  - White spaces
  - Rules and grids
  - Fill color

Delineate rows and cols

- Use white spaces whenever possible
  - Continuity and Proximity
  - Row space to row height < 1:1
- Then use subtle fill colors
- Then use subtle rules
- Avoid grids!

Tables (example)

		Totale					
Squadre	**Punti**	**G**	**V**	**N**	**P**	**F**	**S**
Juventus	38	16	12	2	2	33	10
Inter	34	16	11	1	4	29	17
Napoli	33	16	10	3	3	29	14
Lazio	30	16	9	3	4	24	18
Roma	29	16	9	2	5	38	26
Fiorentina	29	16	8	5	3	29	18
Milan	24	16	7	3	6	28	21
Catania	22	16	6	4	6	22	24
Udinese	22	16	5	7	4	24	24
Atalanta **	21	16	7	2	7	17	23
Parma	20	16	5	5	6	19	22
Chievo	18	16	5	3	8	19	27
Sampdoria *	17	16	5	3	8	19	23
Cagliari	16	16	4	4	8	14	26
Bologna	15	16	4	3	9	17	20
Torino *	15	16	3	7	6	17	21
Palermo	14	16	3	5	8	14	23
Pescara	14	16	4	2	10	12	30
Genoa	12	16	3	3	10	16	28
Siena ******	11	16	4	5	7	15	20

Tables (example)

		Totale					
Squadre	**Punti**	**G**	**V**	**N**	**P**	**F**	**S**
Juventus	38	16	12	2	2	33	10
Inter	34	16	11	1	4	29	17
Napoli	33	16	10	3	3	29	14
Lazio	30	16	9	3	4	24	18
Roma	29	16	9	2	5	38	26
Fiorentina	29	16	8	5	3	29	18
Milan	24	16	7	3	6	28	21
Catania	22	16	6	4	6	22	24
Udinese	22	16	5	7	4	24	24
Atalanta **	21	16	7	2	7	17	23
Parma	20	16	5	5	6	19	22
Chievo	18	16	5	3	8	19	27
Sampdoria *	17	16	5	3	8	19	23
Cagliari	16	16	4	4	8	14	26
Bologna	15	16	4	3	9	17	20
Torino *	15	16	3	7	6	17	21
Palermo	14	16	3	5	8	14	23
Pescara	14	16	4	2	10	12	30
Genoa	12	16	3	3	10	16	28
Siena ******	11	16	4	5	7	15	20

Tables (example)

		Totale					
Squadre	**Punti**	**G**	**V**	**N**	**P**	**F**	**S**
Juventus	38	16	12	2	2	33	10
Inter	34	16	11	1	4	29	17
Napoli	33	16	10	3	3	29	14
Lazio	30	16	9	3	4	24	18
Roma	29	16	9	2	5	38	26
Fiorentina	29	16	8	5	3	29	18
Milan	24	16	7	3	6	28	21
Catania	22	16	6	4	6	22	24
Udinese	22	16	5	7	4	24	24
Atalanta **	21	16	7	2	7	17	23
Parma	20	16	5	5	6	19	22
Chievo	18	16	5	3	8	19	27
Sampdoria *	17	16	5	3	8	19	23
Cagliari	16	16	4	4	8	14	26
Bologna	15	16	4	3	9	17	20
Torino *	15	16	3	7	6	17	21
Palermo	14	16	3	5	8	14	23
Pescara	14	16	4	2	10	12	30
Genoa	12	16	3	3	10	16	28
Siena ******	11	16	4	5	7	15	20

Tables (example)

				Totale			
Squadre	**Punti**	**G**	**V**	**N**	**P**	**F**	**S**
Juventus	38	16	12	2	2	33	10
Inter	34	16	11	1	4	29	17
Napoli	33	16	10	3	3	29	14
Lazio	30	16	9	3	4	24	18
Roma	29	16	9	2	5	38	26
Fiorentina	29	16	8	5	3	29	18
Milan	24	16	7	3	6	28	21
Catania	22	16	6	4	6	22	24
Udinese	22	16	5	7	4	24	24
Atalanta **	21	16	7	2	7	17	23
Parma	20	16	5	5	6	19	22

Tables (example)

				Totale			
Squadre	**Punti**	**G**	**V**	**N**	**P**	**F**	**S**
Juventus	38	16	12	2	2	33	10
Inter	34	16	11	1	4	29	17
Napoli	33	16	10	3	3	29	14
Lazio	30	16	9	3	4	24	18
Roma	29	16	9	2	5	38	26
Fiorentina	29	16	8	5	3	29	18
Milan	24	16	7	3	6	28	21
Catania	22	16	6	4	6	22	24
Udinese	22	16	5	7	4	24	24
Atalanta **	21	16	7	2	7	17	23
Parma	20	16	5	5	6	19	22
Chievo	18	16	5	3	8	19	27
Sampdoria *	17	16	5	3	8	19	23
Cagliari	16	16	4	4	8	14	26
Bologna	15	16	4	3	9	17	20
Torino *	15	16	3	7	6	17	21
Palermo	14	16	3	5	8	14	23
Pescara	14	16	4	2	10	12	30
Genoa	12	16	3	3	10	16	28
Siena ******	11	16	4	5	7	15	20

Arrange Data

- Columns vs. Rows
  - Categorical across columns if few
  - Time-series horizontally across columns
  - Ranked subdivisions down the rows
- Groups and break
  - Use just enough vertical space at the beginning of each group
  - Consistent structure from group to group

Arrange Data

- Column sequence
  - Categorical level on the left
    - Hierarchies from left to right
  - Derived values should be close to the source values
  - Set of quantitative values intended to be compared should be as close as possible
- Sorting
  - Alphabetic order is useful for lookup only
  - Sort according to some inherent order

Formatting text

- Horizontal left to right is the natural orientation
- Alignment
  - Numbers to the right with aligned decimal
  - Dates to the left using same width format
  - Text to the left
  - Center when header is significantly larger

Formatting text

- Date and number formats
  - Include no unnecessary information
    - The level of precision should not exceed the level needed for you communication goal
  - Use data format most familiar to the intended readers
  - Always align data
  - Use thousands separators
  - Whenever possible truncate to multiple of thousands

Formatting text

- Font should be legible
  - Use the same font throughout
  - Avoid **fonts** *with* poor legibility
- Emphasis and color
  - Highlight values
  - Group related items

Summarize values

- Summary values
  - Sum
  - Average
    - Mean
    - Median
  - Occurrence
  - Distribution
    - Range
    - Std Deviation

GRAPH DESIGN

Graph design principles

- Encode quantities to correspond accurately to the visual scale
  - Distance between tick marks must correspond to the values they represent
  - Include the zero value in quantitative scales
    - Warn when its not
- Avoid 3D display of quantitative data
  - Data-ink ratio!

Graph components

- Data components
- Support components

Points

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

Bars

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

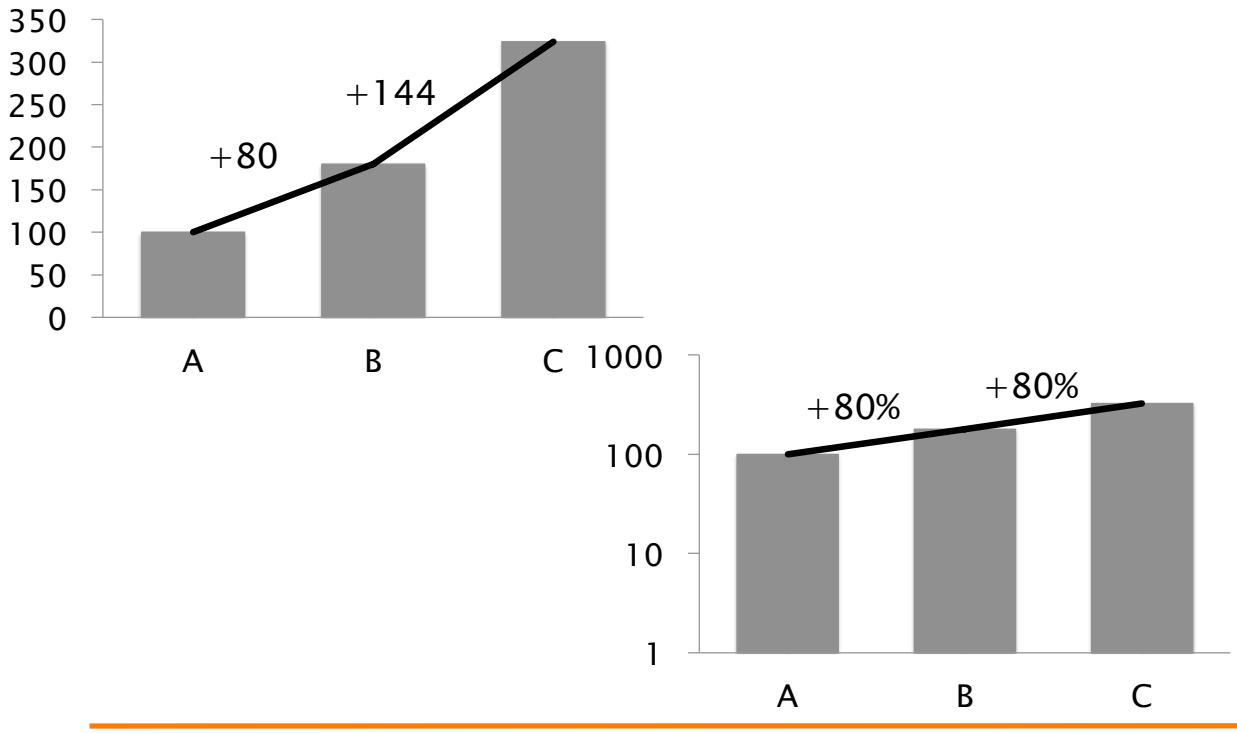
Bars

- Fill
  - Avoid fill patterns
  - Colors must be clearly distinct
  - Balanced colors for similar importance data
  - Intense colors to highlight important values
- Borders when
  - Fill colors not clearly distinct from background
    - Light gray may work
  - Highlight a single bar
- Bars must start at zero
  - Except when representing ranges

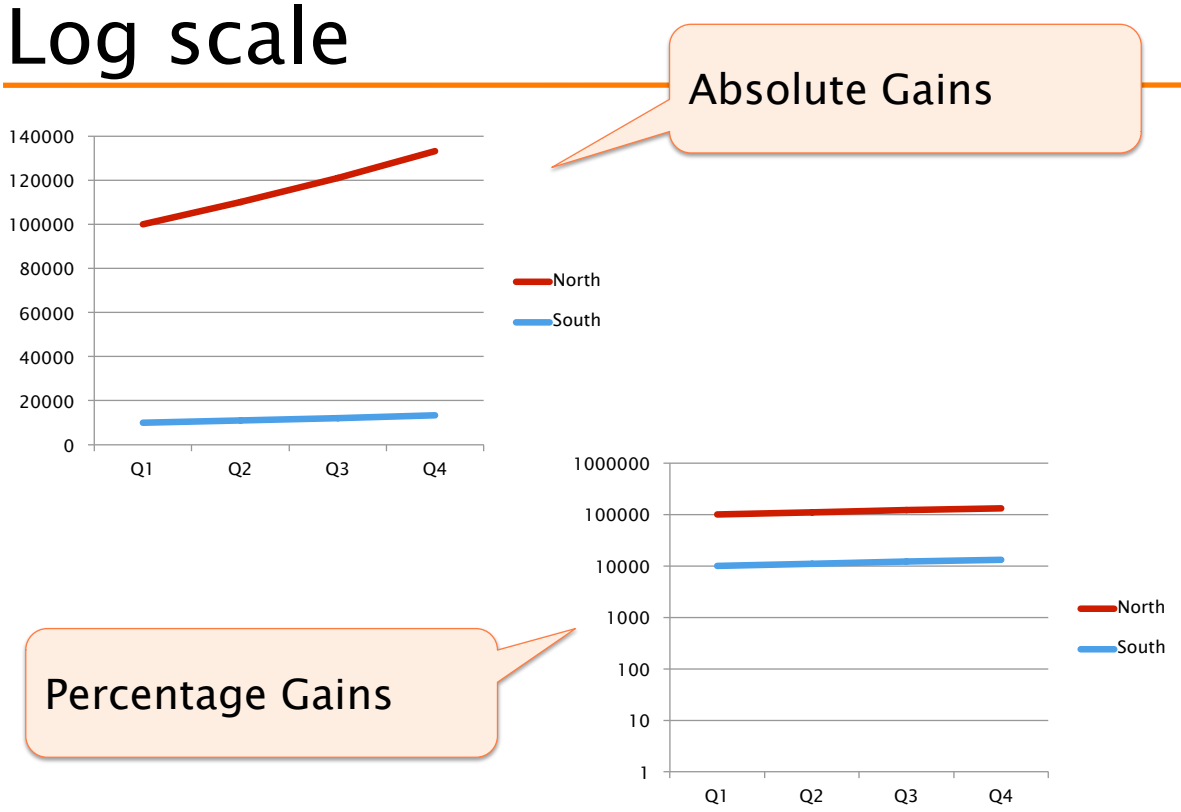
Log scale

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

Log scale



Log scale



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

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

Graph area

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

Other

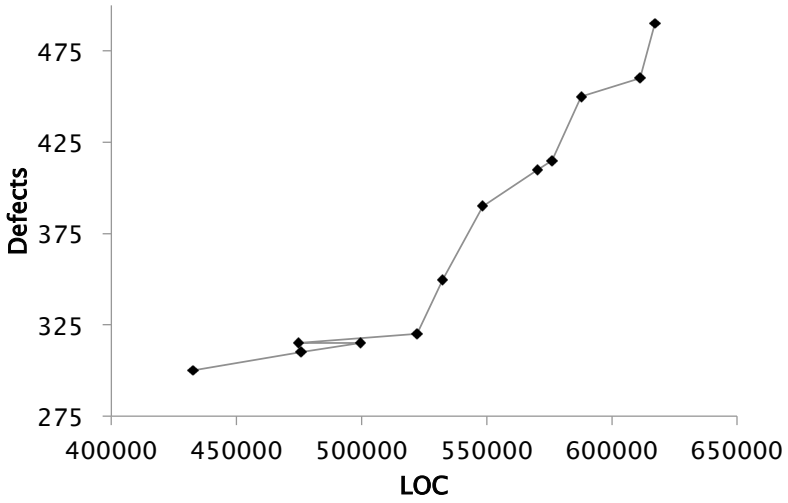
- Text should be as close as possible to the object it complements
- Number of categorical subdivisions
  - Perceptual limit is between 5 and 8
  - Limit is independent of the visual attribute used to encode it
- Number of axis should be 2
  - 1 is fine for horizontal bars
    - continuity gestalt principle

Multiple variable graph

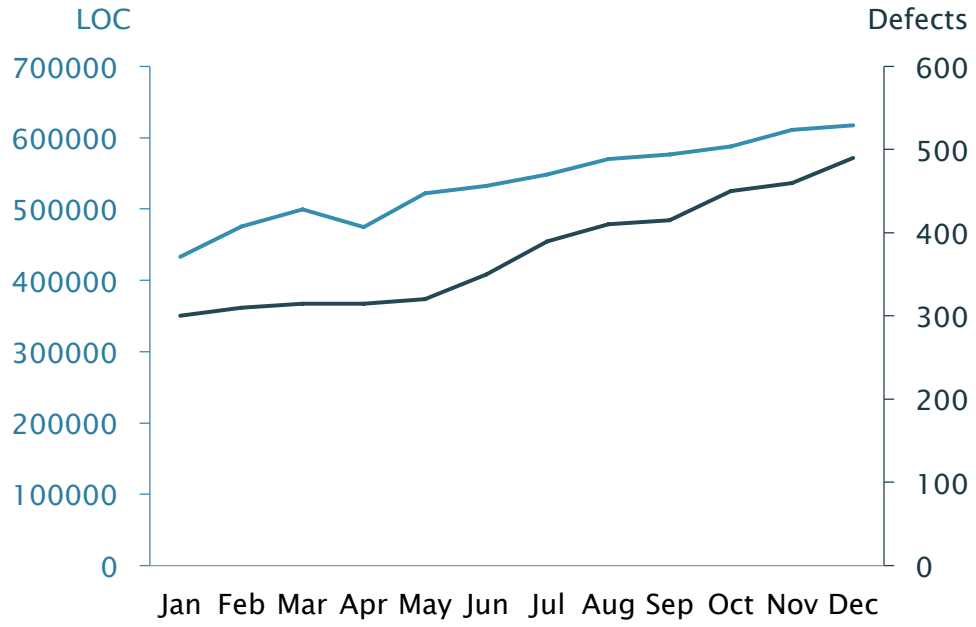
- Multiple unit of measure
  - Double quantitative (y) axis
- Multiple graphs in a series

Multiple variable graph

- Example
  - Defects and LOCs for each month

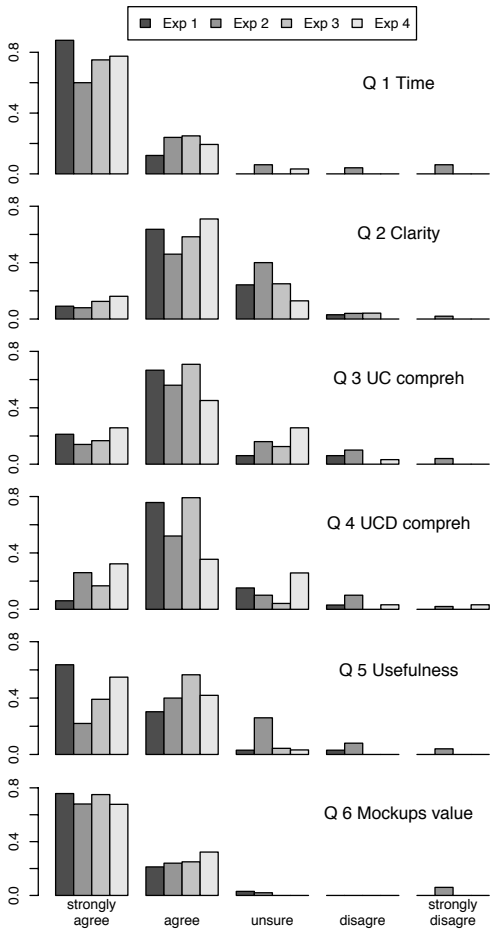


Double scale



Series of graph

- Consistency
  - Same scale
  - Same categorical levels
  - Same ordering of categorical levels
- Arrangement
  - Align axis that involve comparison
    - Possibly along a matrix



Series of graphs

- 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

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.