

VISUALIZATION BASICS

(A RECAP)



UNIVERSITY OF
CALGARY



<https://tinyurl.com/DATA605-W2020>

WHAT IS DATA VISUALIZATION?

**“THE USE OF COMPUTER-
GENERATED, INTERACTIVE,
VISUAL REPRESENTATIONS
OF DATA TO AMPLIFY
COGNITION.”**



CARD



MACKINLAY



SHNEIDERMAN

**INFORMATION VISUALIZATION:
USING VISION TO THINK (1999)**

WHY VISUALIZE DATA?

RECORD INFORMATION

BLUEPRINTS, PHOTOGRAPHS, SEISMOGRAPHS, ETC. ...

ANALYZE DATA TO SUPPORT REASONING

- DEVELOP AND ASSESS HYPOTHESES
- DISCOVER ERRORS IN DATA
- EXPAND MEMORY
- FIND PATTERNS

COMMUNICATE INFORMATION TO OTHERS

SHARE AND PERSUADE
COLLABORATE AND REVISE

PEOPLE HAVE BEEN
VISUALLY AND PHYSICALLY
REPRESENTING DATA
FOR A LONG TIME

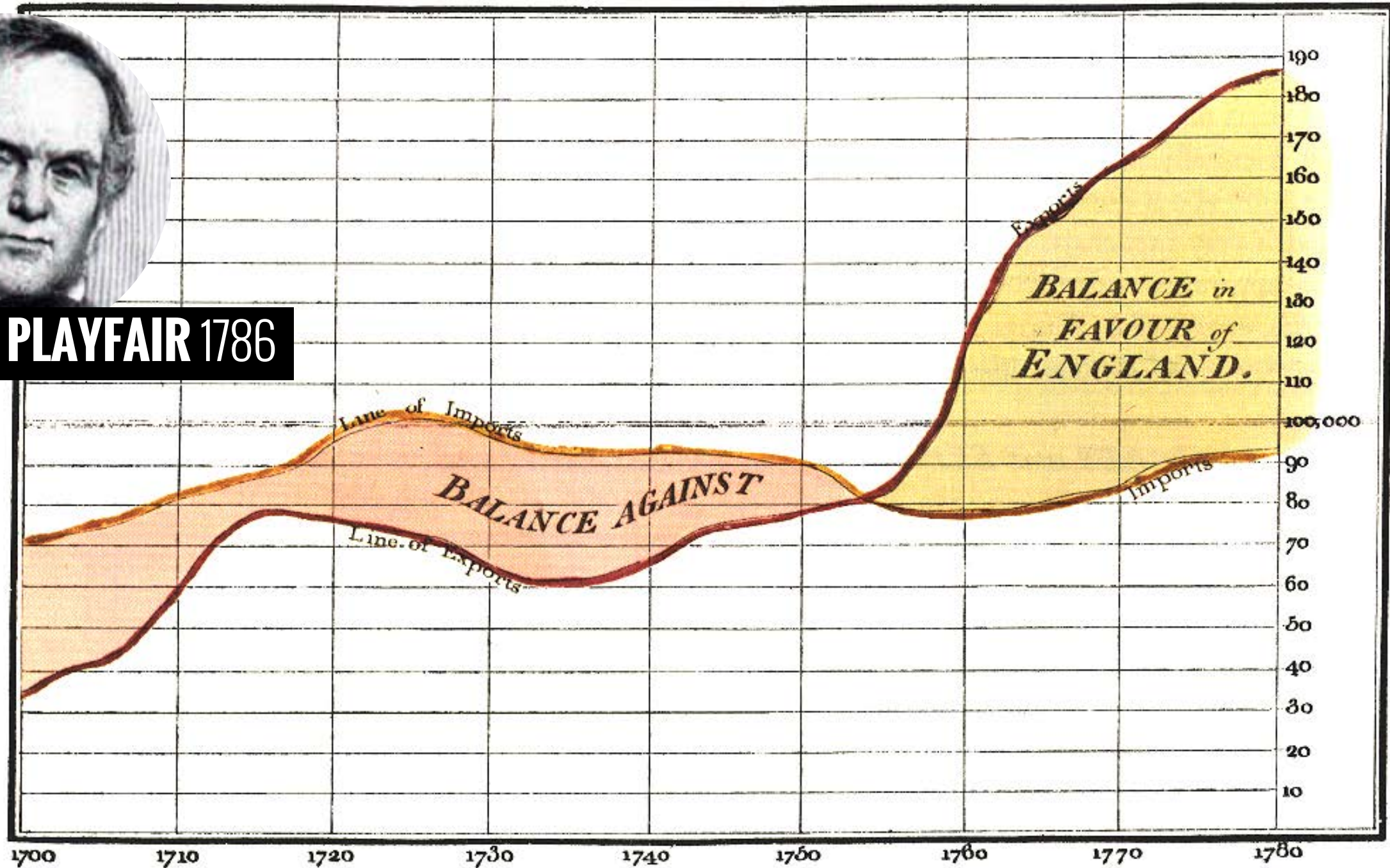


MESOPOTAMIAN CLAY TOKENS 5500 BCE

Exports and Imports to and from DENMARK & NORWAY from 1700 to 1780.



WILLIAM PLAYFAIR 1786



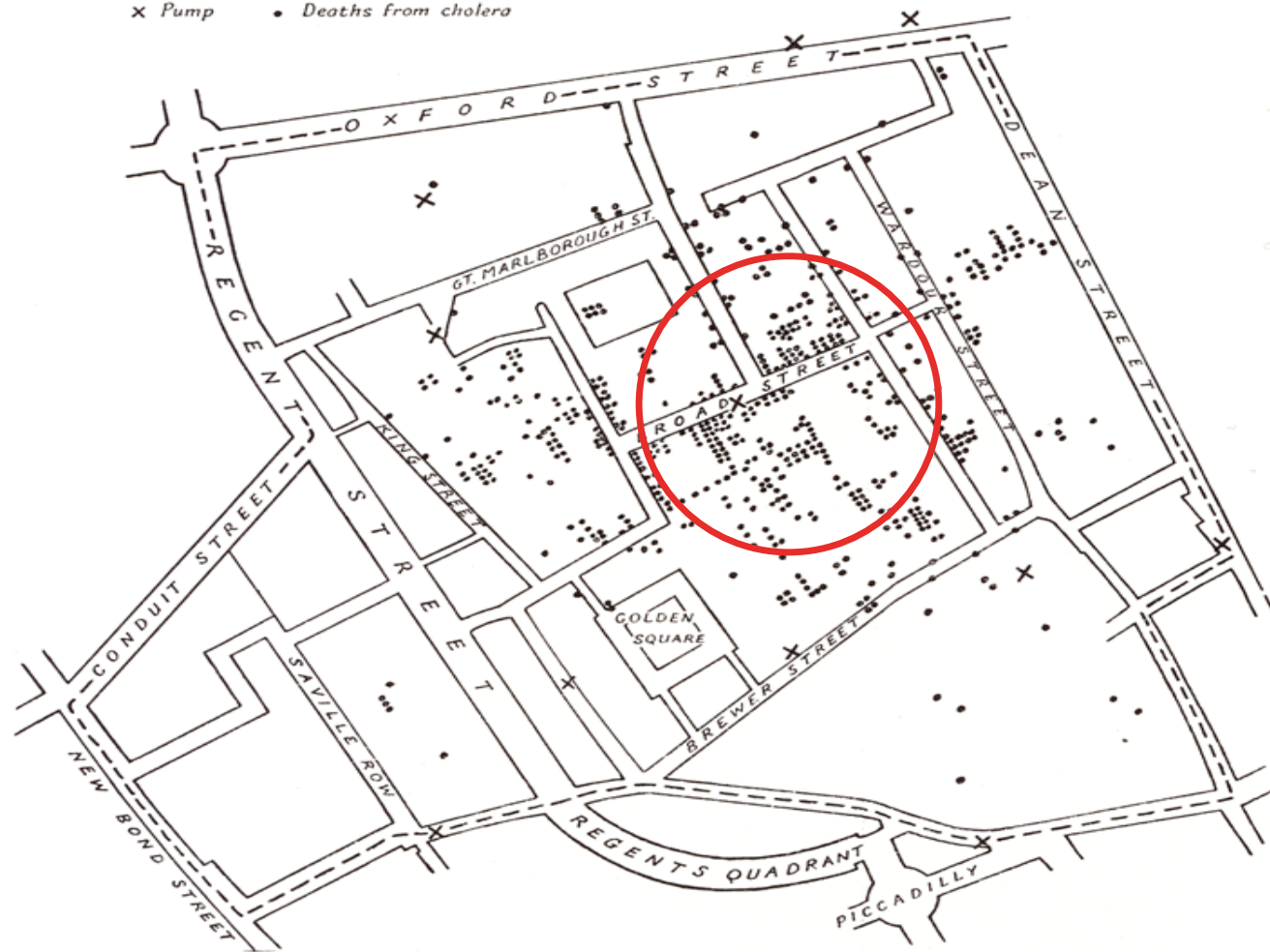


JOHN SNOW 1854

IN 1854 SNOW PLOTTED THE POSITION OF EACH CHOLERA CASE ON A MAP. [FROM TUFTE 83]

50 0 50 100 150 200
Yards

X Pump • Deaths from cholera

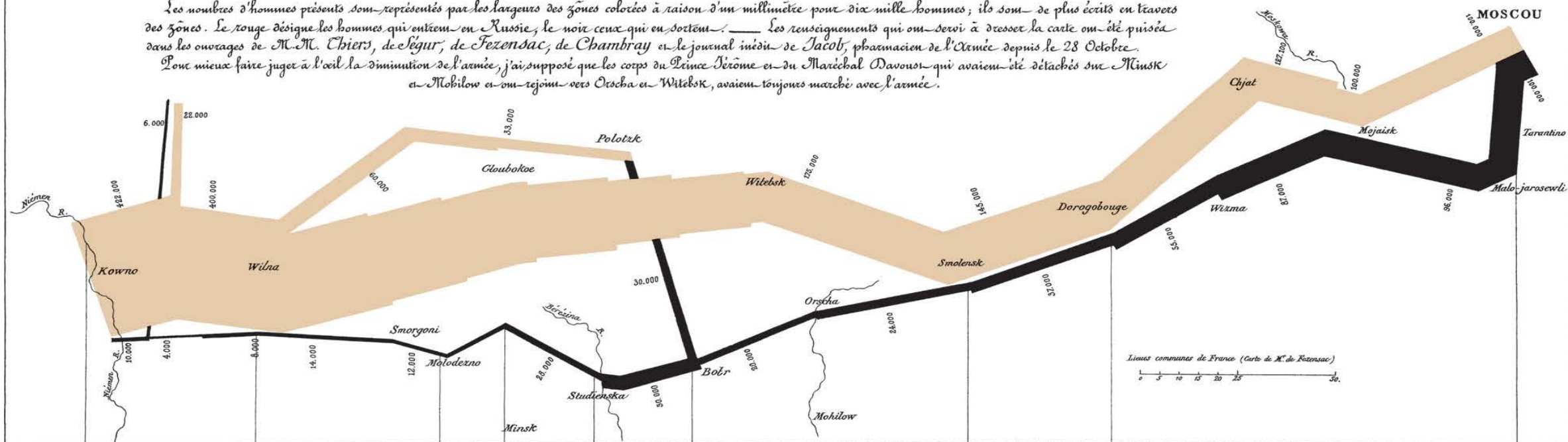


Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813.

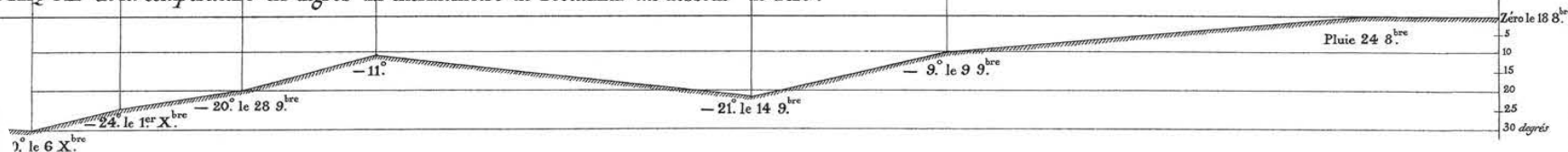
Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite Paris, le 20 Novembre 1869.

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 Ségur, de Fezensac, de Chambray et le journal intime 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 et ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l'armée.



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



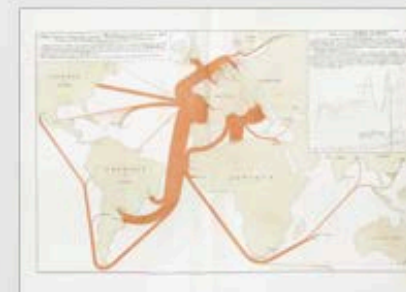
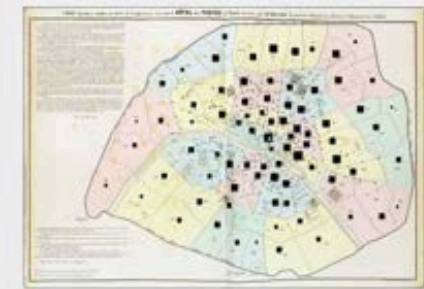
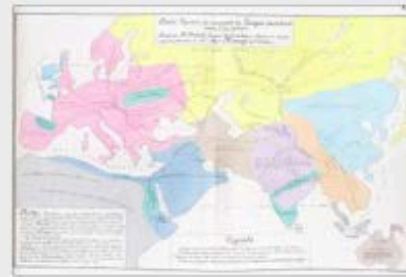
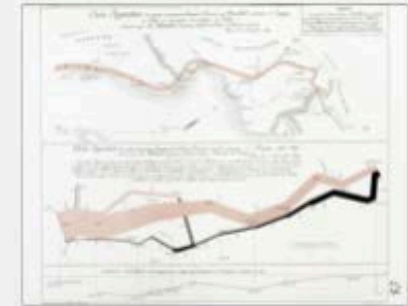
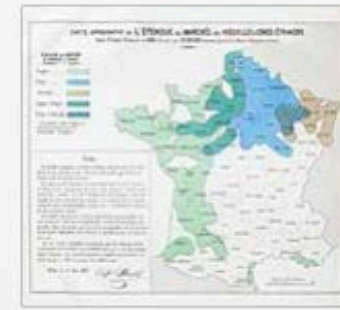
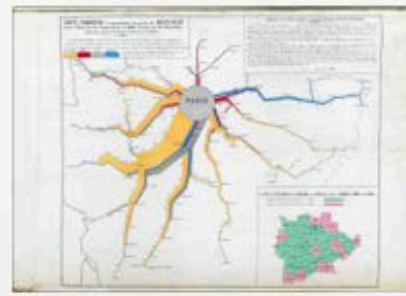
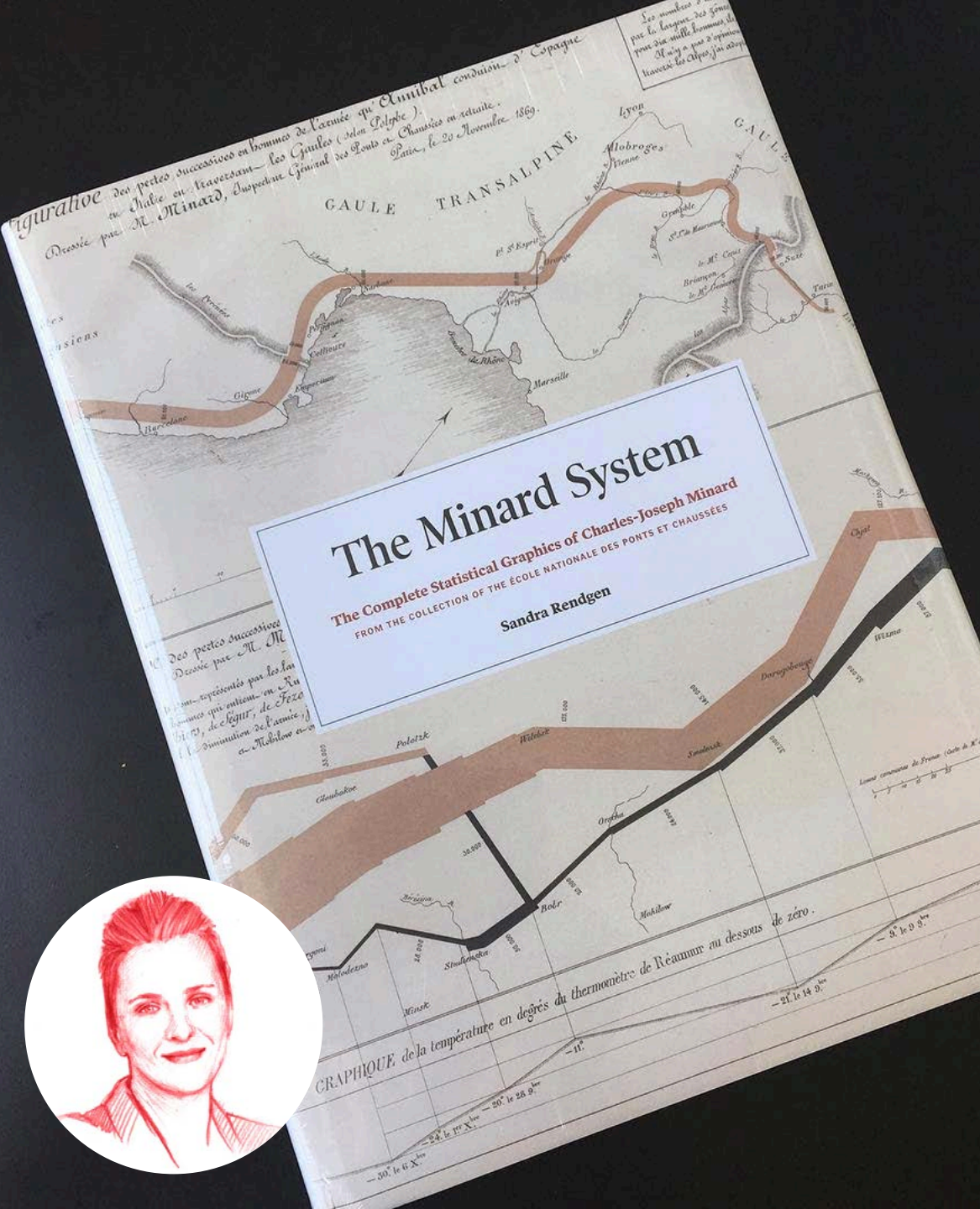
Les Cosaques passent le Niemen gelé.

Autog. par Ragnier, S. Par.



Imp. Lith. Ragnier et Doucet.

CHARLES JOSEPH MINARD 1869





JACQUES BERTIN

VISUAL LANGUAGE IS A SIGN SYSTEM

Images perceived as a set of signs

Sender encodes information in signs

Receiver decodes information from signs

SÉMIOLOGIE GRAPHIQUE (THE SEMIOLOGY OF GRAPHICS) 1967

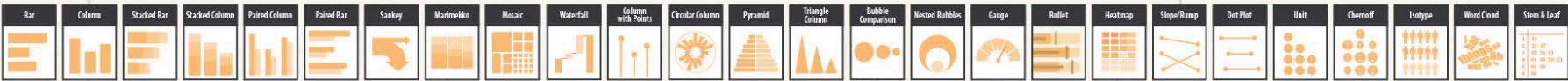
The Graphic Continuum shows several ways that data can be illustrated individually or combined to show relationships. Use of various shapes, chart types, and colors can help identify patterns, tell stories, and reveal relationships between different sets and types of data. Bar charts, or histograms, for example, can illustrate a distribution of data over time, but they also can show categorical or geographic differences. Scatterplots can illustrate data from a single instance or for a period, but they also can be used to identify a distribution around a mean.

This set of charts does not constitute an exhaustive list, nor do the connections represent every possible pathway for linking data and ideas. Instead, the Graphic Continuum identifies some presentation methods, and it illustrates some of the connections that can bind different representations together. The six groups do not define all possibilities: Many other useful, overlapping data types and visualization techniques are possible.

This chart can guide graphic choices, but your imagination can lead the way to other effective ways to present data.

COMPARING CATEGORIES

Compare values across categories



GEOSPATIAL

Relate data to its geography



PART-TO-WHOLE

Visualizations that relate the part of a variable to its total



RELATIONSHIP

Illustrates correlations or relationships between variables



TIME

Track changes over time



A Fan Chart combines a line chart for observed past data with a range for possible future values

A Slope or Bump Chart connects categories across vertical axes instead of showing all points

A Connected Scatterplot illustrates changes in correlations over time

A Heatmap uses color to show high-frequency comparisons. A Correlation Matrix uses the size of shapes to illustrate correlations

An Arc-Time Chart presents connections across time. The Arc-Connection Chart shows connections between observations

A Circle-Packing diagram illustrates a hierarchy. A Dotting Map uses codes without the explicit use of a map to show a geographic distribution

A Bubble Chart encodes data based on circle size. A Bubble Map can be used to illustrate a geographic distribution

A Map with Columns encodes data within a geographic frame of reference

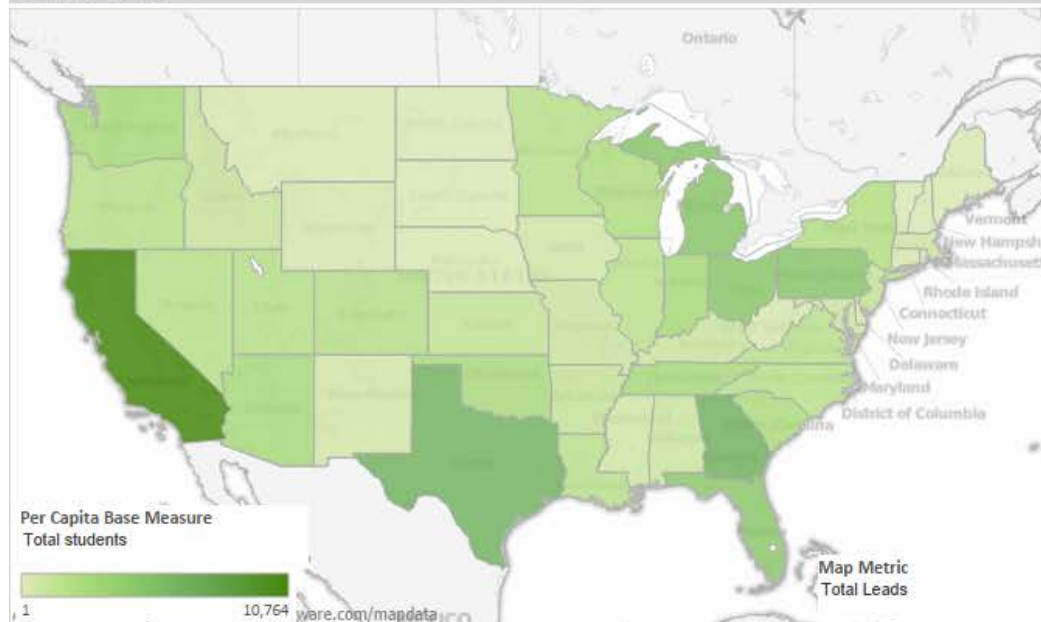
A Heatmap uses color to show high-frequency comparisons. A Treemap uses rectangles to show part-to-whole relationships

ENCODING

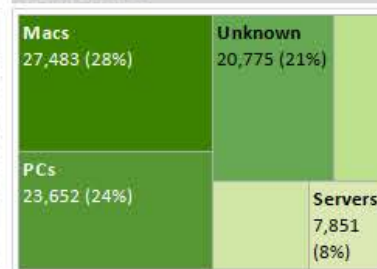
DATA → MARKS

**TO EFFICIENTLY LEVERAGE
HUMAN PERCEPTION**

Leads by State

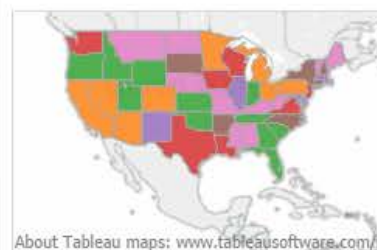


Product Area



Show Classifications Of

State Type



Response Time

Response Time	Convert %	Leads	Converted
< 2 Hrs	6.46%	5,310	343
< 1 Day	4.67%	9,556	446
Later	3.89%	84,134	3,270

Lead Volume Change

	Leads	2012	2013	WoW Change	2012	2013	YoY Change	2012	2013
1	4,475	1,933					-57%		
2	3,249	1,645		-27%	-15%		-49%		
3	1,714	2,035		-47%	24%		19%		
4	1,322	4,854		-23%	139%		267%		
5	1,476	2,743		12%	-43%		86%		
6	5,300	2,643		259%	-4%		-50%		
7	3,624	2,420		-32%	-8%		-33%		
8	360	1,888		-90%	-22%		424%		
9		1,051		-100%	-44%				
10		1,113			6%				
11	1,196	2,639			137%		121%		
12	4,418	2,345		269%	-11%		-47%		
13	3,990	2,904		-10%	24%		-27%		
14	1,155	2,358		-71%	-19%		104%		
15		1,809		-100%	-23%				
16		1,086			-40%				
17		1,193			10%				
18		2,941			147%				
19		2,889			-2%				
20		2,616			-9%				
21		3,358			28%				
22		2,554			-24%				
23		1,188			-53%				
24		1,326			12%				
25		2,515			90%				
26		2,411			-4%				
27		2,166			-10%				
28		2,494			15%				
29		1,742			-30%				

Summary

Lead Gen Budget	\$3,226,785
Leads	99,000
Budget per Lead	\$32.59
Converted	4,059
Budget per Conversion	\$794.97
Convert %	4.10%

Filters

Region

State Type

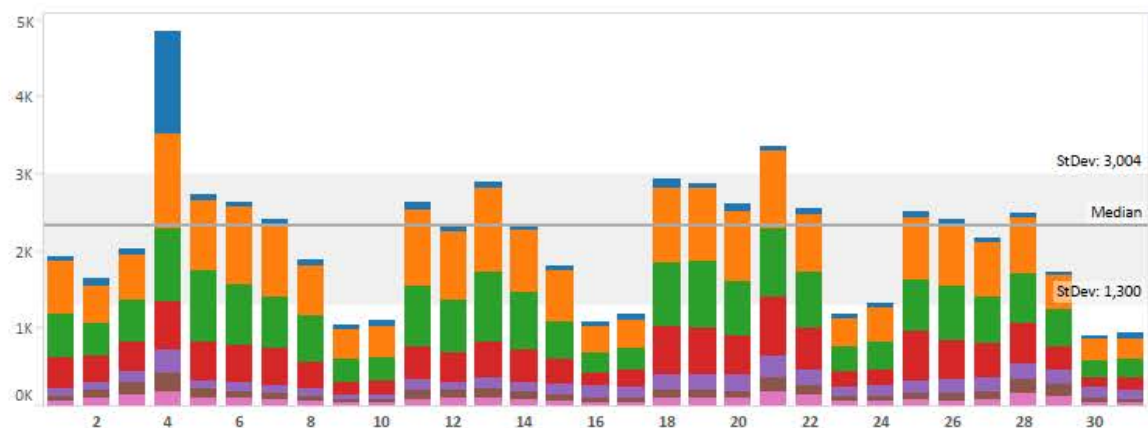
Lead Source

All

Generated By

All

Breakdown Lead By Day



Measure: Lead

Date Interval: Day

Color By: State Type

Unspecified

Type A

Type B

Type C

Type D

Type E

Type F

A CONTEMPORARY UNDERSTANDING

PERCEPTION (How we see information.)

COGNITION (How we process it.)

INTERACTION (How we connect on it)

How should we **represent** data to best achieve a particular **task or goal**?

DATA

Different kinds of data have different characteristics.

Not all can be analyzed or represented in the same ways.

RELATIONAL DATA MODEL

Represent data as a **TABLE** (*relation*)

ROW (*tuple*) represents a single record
Each record is a fixed-length tuple

COLUMN (*attribute*) represents a single *variable*
Each has a name and a data type

SCHEMA – table's set of names and data types

DATABASE - a collection of tables

Month	Treatment	Pressure
March	Control	165
March	Placebo	163
March	300 mg	166
March	450 mg	168
April	Control	162
April	Placebo	159
April	300 mg	161
April	450 mg	163
May	Control	164

Blood Pressure Study (4 treatments, 6 months)



S. S. STEVENS

NOMINAL, ORDINAL, AND QUANTITATIVE

N – NOMINAL (or “CATEGORICAL”) – LABELS

- Fruits: Apples, oranges, ...

O – ORDINAL – HAS A MEANINGFUL ORDERING

- Quality of meat: Grade A, AA, AAA
- Can be ordered and counted, but not measured

Q – QUANTITATIVE

- INTERVAL
 - No clear zero (or arbitrary)
 - E.g. dates, latitude, longitude
 - Usually compare differences (“intervals”)
- RATIO
 - Meaningful origin (often zero)
 - E.g. Physical measurement (Length, Mass, ...)
 - Counts and amounts

NOMINAL, ORDINAL, AND QUANTITATIVE

N - Nominal (labels)

- Operations: $=$, \neq

O - Ordered

- Operations: $=$, \neq , $<$, $>$

Q - Interval (Location of zero arbitrary)

- Operations: $=$, \neq , $<$, $>$, $-$
- Can measure distances or spans

Q - Ratio (zero fixed)

- Operations: $=$, \neq , $<$, $>$, $-$, \div
- Can measure ratios or proportions

DIMENSIONS AND MEASURES

DIMENSIONS

Discrete variables describing data

Categories, Names, Dates, etc. (independent vars)

MEASURES

Data values that can be aggregated

Numbers to be analyzed (dependent vars)

- Aggregate as sum, count, average, std. deviation, etc.

Month	Treatment	Pressure
March	Control	165
March	Placebo	163
March	300 mg	166
March	450 mg	168
April	Control	162
April	Placebo	159
April	300 mg	161
April	450 mg	163
May	Control	164

Blood Pressure Study (4 treatments, 6 months)

EXAMPLE: CENSUS DATA

Year: 1901- 2011 (every 5 years)

Age: 0 - 90+

Sex: Male, Female

Marital Status: Single, Married, Divorced, ...

People: # of people in group

NOMINAL, ORDINAL, OR QUANTITATIVE?

Year

Q-Interval *(0)*

Age

Q-Ratio *(0)*

Sex

N

Marital Status

N

People

Q-Ratio

DIMENSION OR MEASURE?

Year

Age

Sex

Marital Status

People

Dimension

Depends!

Dimension

Dimension

Measure

WHY IS THIS IMPORTANT?

Nominal, Ordinal, and Quantitative data are best encoded in different ways.

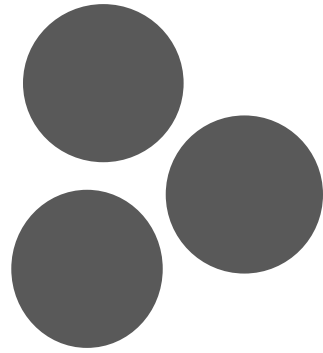
Dimensions and Measures are important concepts in many analysis tools.

ENCODING

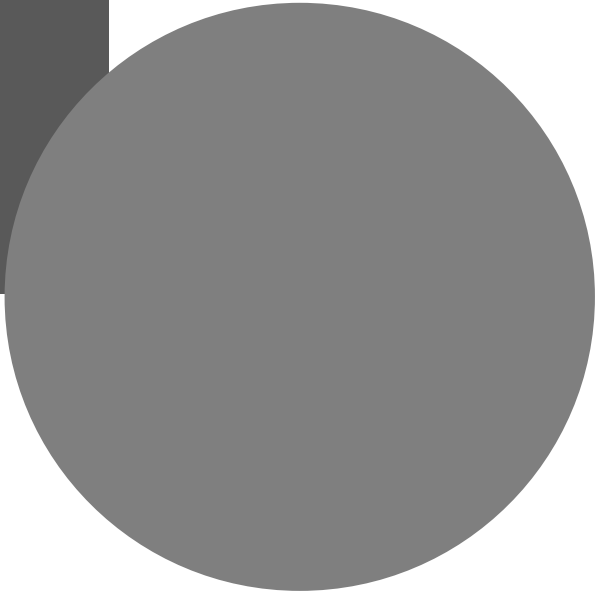
DATA → **MARKS**

(In ways that make it easier to make comparisons, see trends, and answer questions.)

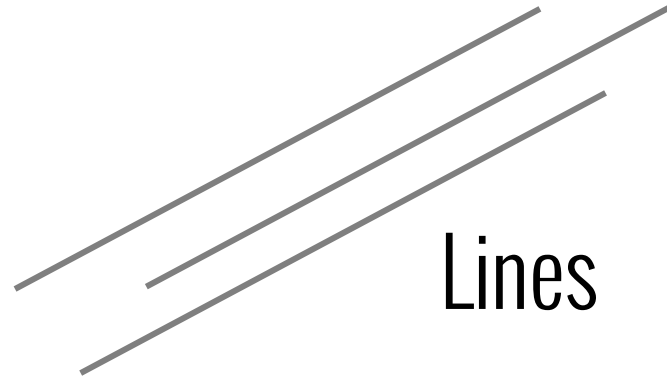
MARKS



Points



Areas



Lines

SEMIOLOGY OF GRAPHICS [BERTIN, 1967]



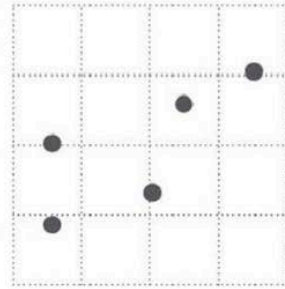
VISUAL VARIABLES

Position (2D)
Size
Value
Texture
Color
Orientation
Shape

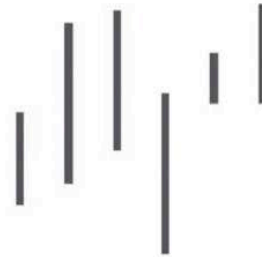
LES VARIABLES DE L'IMAGE									
POINTS			LIGNES			ZONES			
XY 2 DIMENSIONS DU PLAN									
Z TAILLE									
VALEUR									
LES VARIABLES DE SÉPARATION DES IMAGES									
GRAIN									
COULEUR									
ORIENTATION									
FORME									

Position

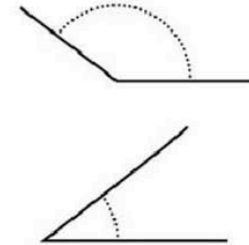
Where in space the data is

**Length**

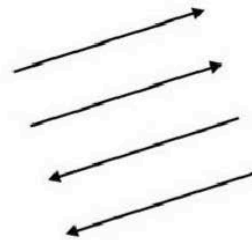
How long the shapes are

**Angle**

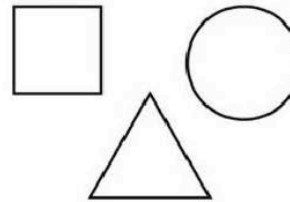
Rotation between vectors

**Direction**

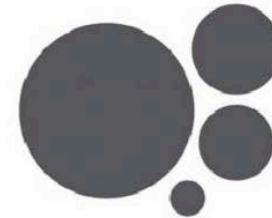
Slope of a vector in space

**Shapes**

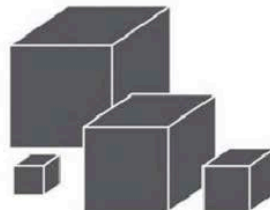
Symbols as categories

**Area**

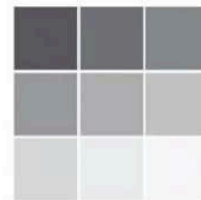
How much 2-D space

**Volume**

How much 3-D space

**Color saturation**

Intensity of a color hue

**Color hue**

Usually referred to as color



FIGURE 3-3 *Visual cues*

BERTIN'S “LEVELS OF ORGANIZATION”

Position

N	O	Q
----------	----------	----------

Size

N	O	Q
----------	----------	----------

Value

N	O	q
----------	----------	----------

Texture

N	o	
----------	----------	--

Color

N		
----------	--	--

Orientation

N		
----------	--	--

Shape

N		
----------	--	--

Nominal

Ordered

Quantitative

Note: $Q < O < N$

Note: Bertin also breaks visual variables down into differentiating (\neq) and associating (\equiv)

GOOD ENCODINGS

More Accurate



Less Accurate

Quantitative		Ordinal		Nominal	
Position		Position		Position	
Length		Density		Hue	
Angle		Saturation		Density	
Slope		Hue		Saturation	
Area		Length		Shape	
Density		Angle		Length	
Saturation		Slope		Angle	
Hue		Area		Slope	
Shape		Shape		Area	

[JACQUES BERTIN REFINED BY CLEVELAND & MCGILL THEN BY CARD & MACKINLAY]

➔ **Magnitude Channels: Ordered Attributes**

Position on common scale 


Position on unaligned scale 

Length (1D size) 

Tilt/angle 

Area (2D size) 

Depth (3D position) 

Color luminance 

Color saturation 

Curvature 


Volume (3D size) 

Most

Effectiveness

Least

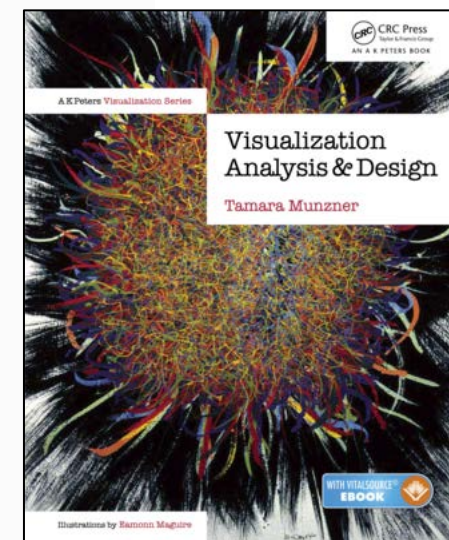
➔ **Identity Channels: Categorical Attributes**

Spatial region 

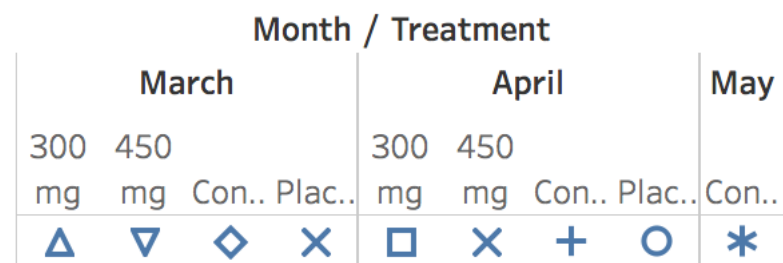
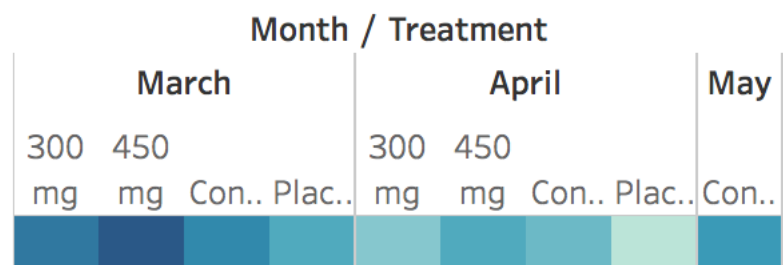
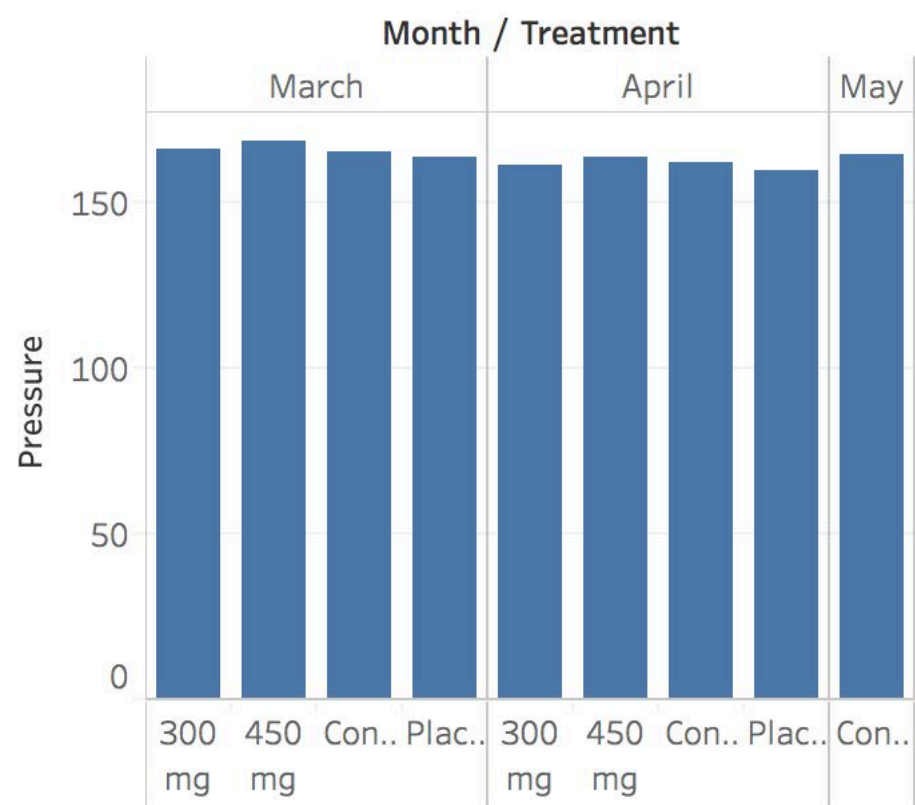
Color hue 

Motion 

Shape 



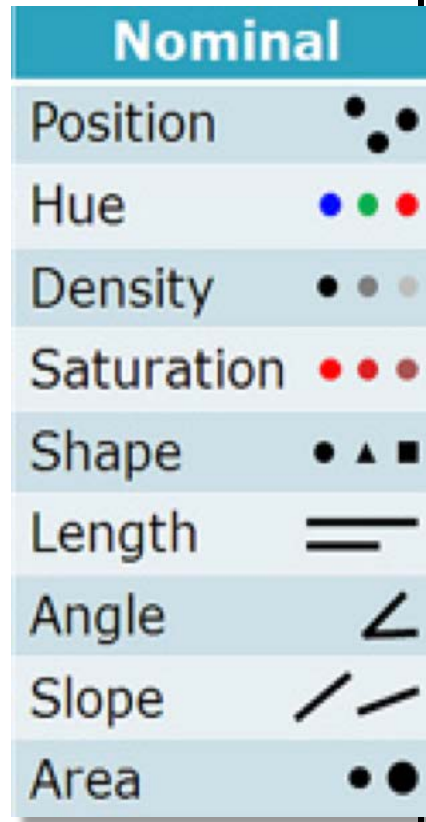
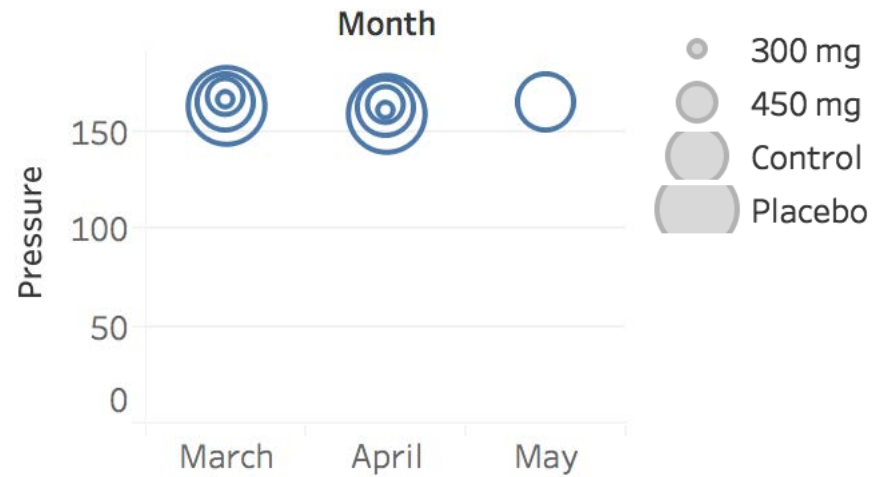
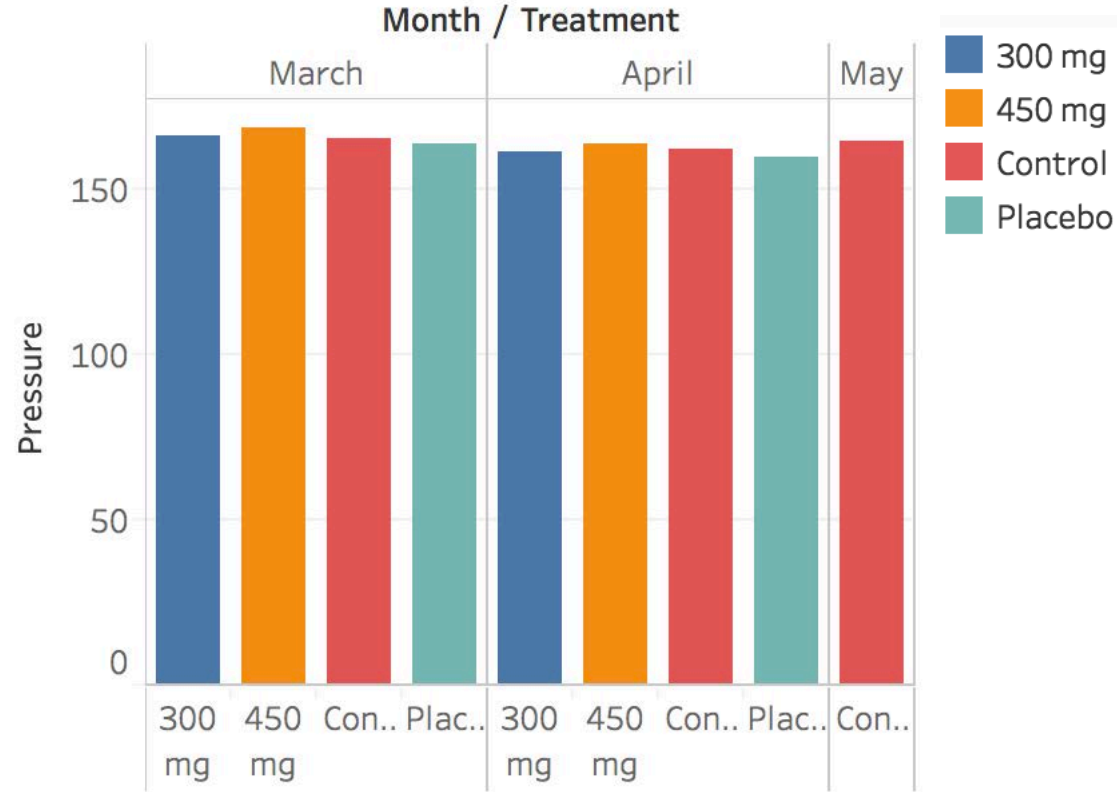
TAMARA MUNZNER



- 159
- 161
- + 162
- × 163
- * 164
- ◇ 165
- △ 166
- ▽ 168

Quantitative	
Position	
Length	
Angle	
Slope	
Area	
Density	
Saturation	
Hue	
Shape	

Month	Treatment	Pressure
March	Control	165
March	Placebo	163
March	300 mg	166
March	450 mg	168
April	Control	162
April	Placebo	159
April	300 mg	161
April	450 mg	163
May	Control	164



Month	Treatment	Pressure
March	Control	165
March	Placebo	163
March	300 mg	166
March	450 mg	168
April	Control	162
April	Placebo	159
April	300 mg	161
April	450 mg	163
May	Control	164

IF YOU REMEMBER ANYTHING...

If you want to distinguish **categories**, use...
position, color, or maybe shape

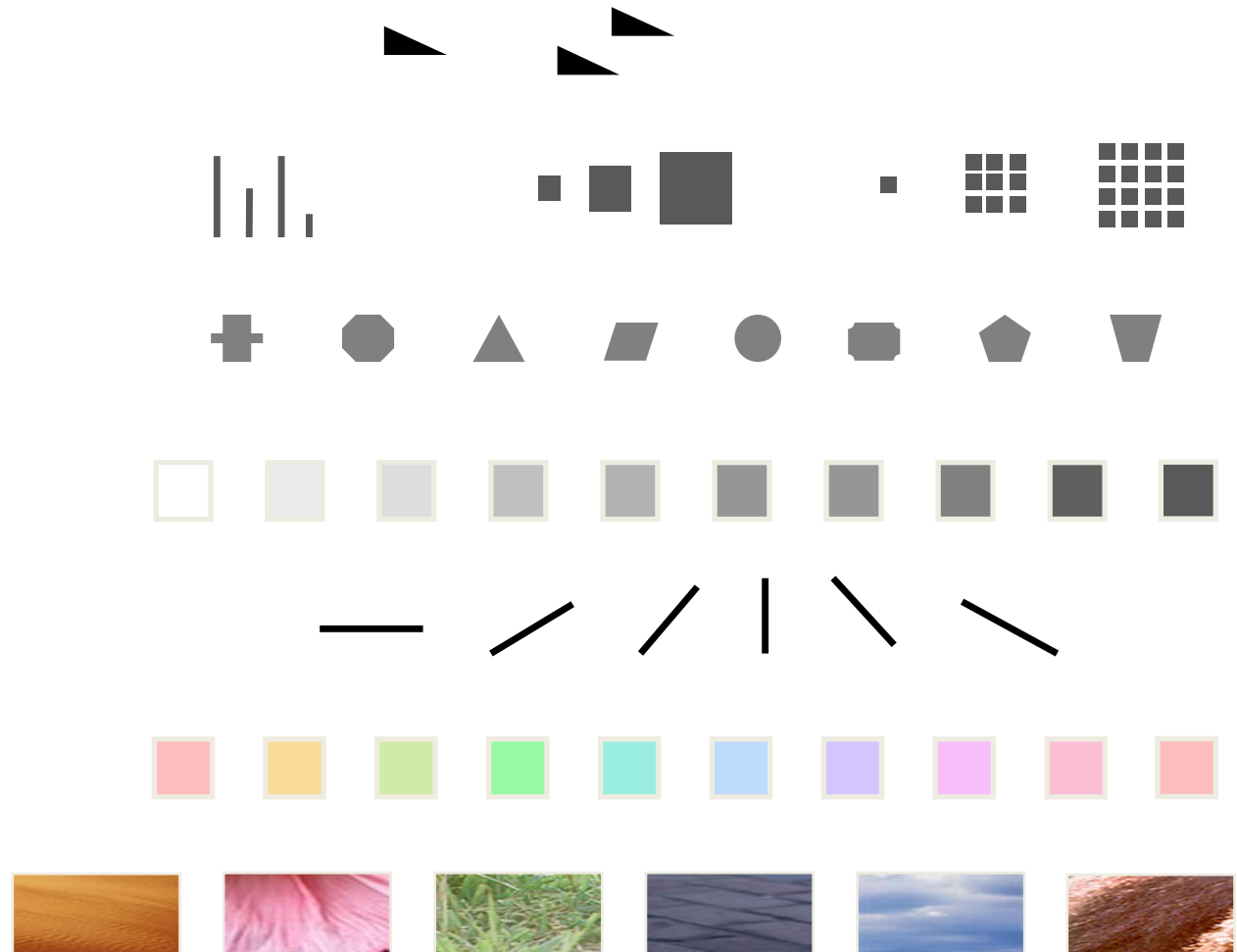
If you want to show numbers use...
position or length, then area, then color

VISUAL VARIABLES

(In more detail.)

VISUAL VARIABLES VARY MARKS

- POSITION
changes in the x, y, (z) location
- SIZE
change in length, area, repetition
- SHAPE
infinite number of shapes
- VALUE
changes from light to dark
- ORIENTATION
changes in alignment
- COLOR
changes in hue at a given value
- TEXTURE
variation in pattern



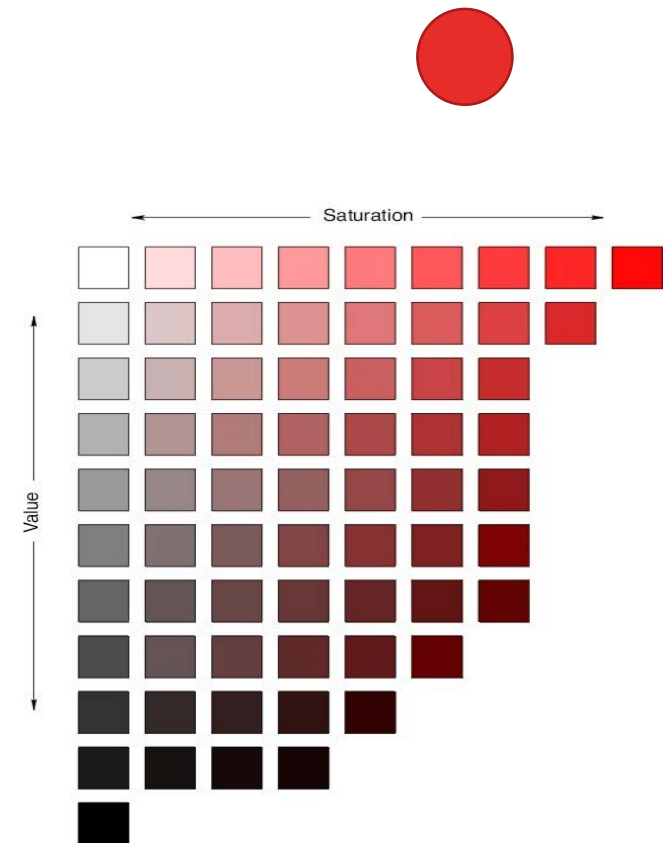
ADDITIONAL VARIABLES...

■ MOTION

Direction, acceleration, speed, frequency, onset, ‘personality’

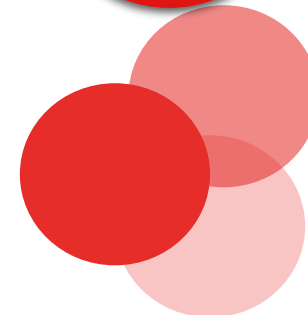
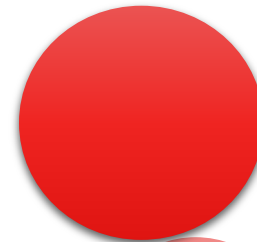
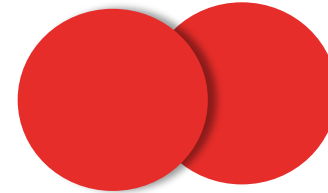
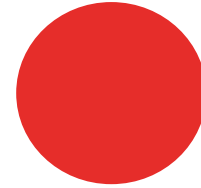
■ COLOR

Bertin uses “color” (**hue**) and “value” (**brightness**) but not **saturation**



ALSO...

- FLICKER
 - frequency, rhythm, appearance
- DEPTH? 'QUASI' 3D
 - depth, occlusion, aerial perspective, binocular disparity
- ILLUMINATION
- TRANSPARENCY



CHARACTERISTICS OF VISUAL VARIABLES

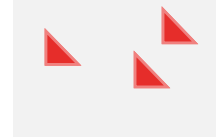
SEMIOLOGY OF GRAPHICS [BERTIN, 1967]

- **SELECTIVE:**
Can this variable allow us to spontaneously differentiate/isolate items from groups?
- **ASSOCIATIVE:**
Can this variable allow us to spontaneously group items in a group?
- **ORDERED:**
Can this variable allow us to spontaneously perceive an order?
- **QUANTITATIVE:**
Is there a numerical reading obtainable from changes in this variable?
- **VARIATIONS (“LENGTH” OR “RESOLUTION”):**
How many different *distinguishable* variations are there?

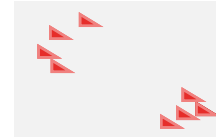
VISUAL VARIABLE: POSITION



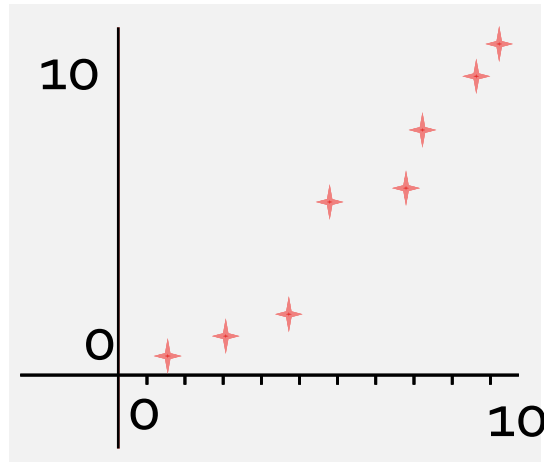
SELECTIVE



ASSOCIATIVE



QUANTITATIVE



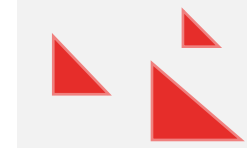
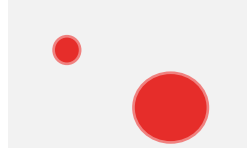
ORDER



VARIATIONS

VISUAL VARIABLE: SIZE

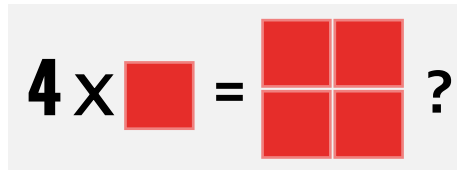
✓ SELECTIVE



✓ ASSOCIATIVE

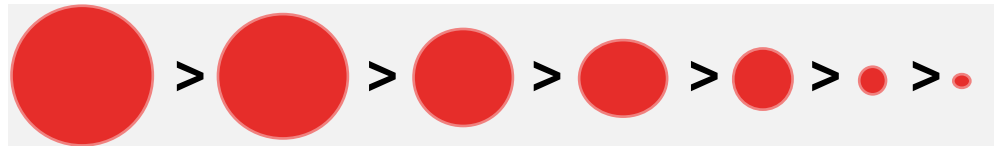


⚡ QUANTITATIVE



✓ ORDER

✓ VARIATIONS

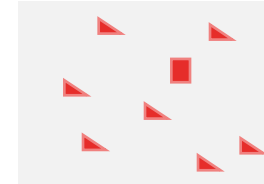
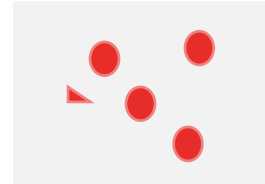
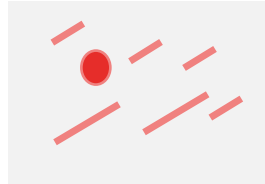


theoretically infinite but practically limited
association and selection ~ 5
and variations ~ 20

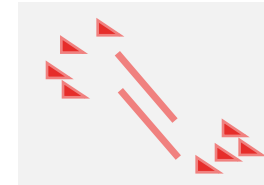
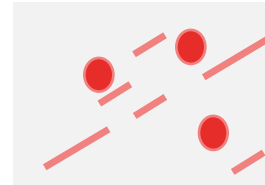
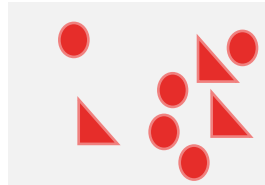
VISUAL VARIABLE: SHAPE



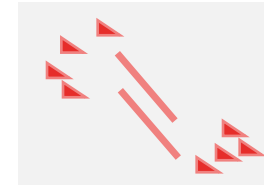
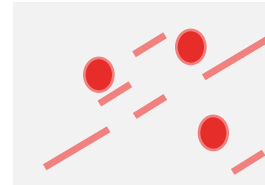
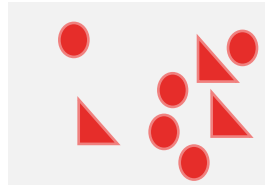
SELECTIVE



ASSOCIATIVE



QUANTITATIVE

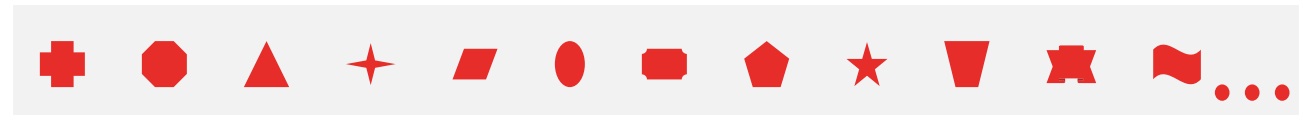


ORDER



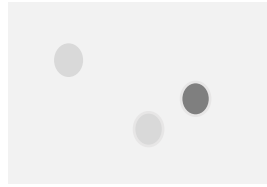
VARIATIONS

basically
infinite!

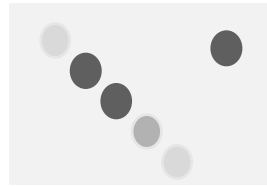


VISUAL VARIABLE: LUMINOSITY

✓ **SELECTIVE**



✓ **ASSOCIATIVE**



≠ **QUANTITATIVE**

✓ **ORDER**

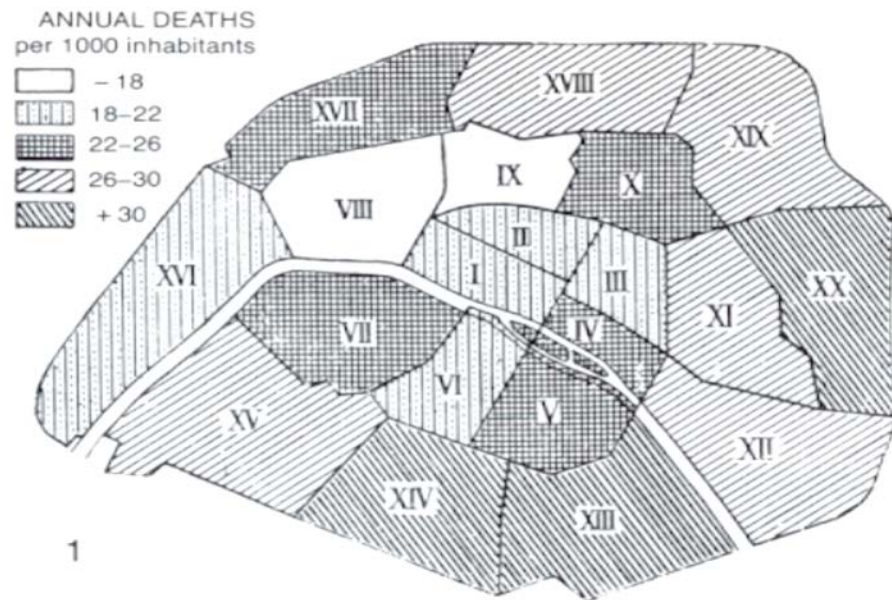


✓ **VARIATIONS**

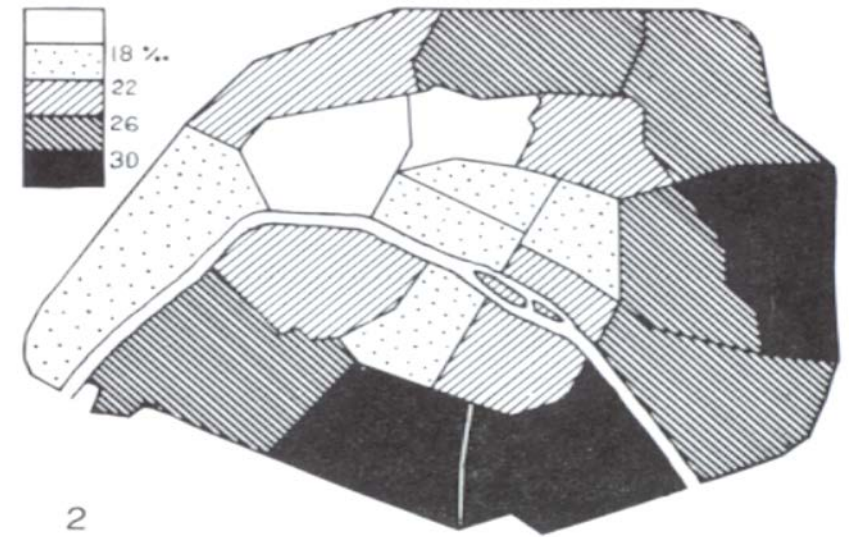
theoretically infinite but practically limited
association and selection <7
and variations ~ 10

VALUE

Perceived as ordered and cannot be reordered



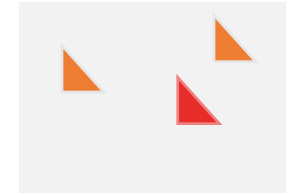
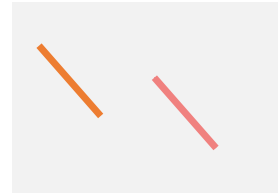
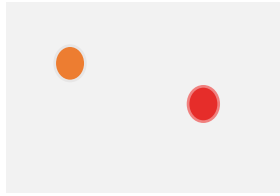
Values **not ordered correctly** according to legend
information **has to be read point by point**



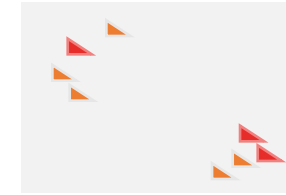
Values **ordered correctly**
image much more useful

VISUAL VARIABLE: COLOR

✓ **SELECTIVE**



✓ **ASSOCIATIVE**



≠ **QUANTITATIVE**

≠ **ORDER**

✓ **VARIATIONS**



theoretically infinite but practically limited
association and selection <7
and variations ~ 10

VISUAL VARIABLES

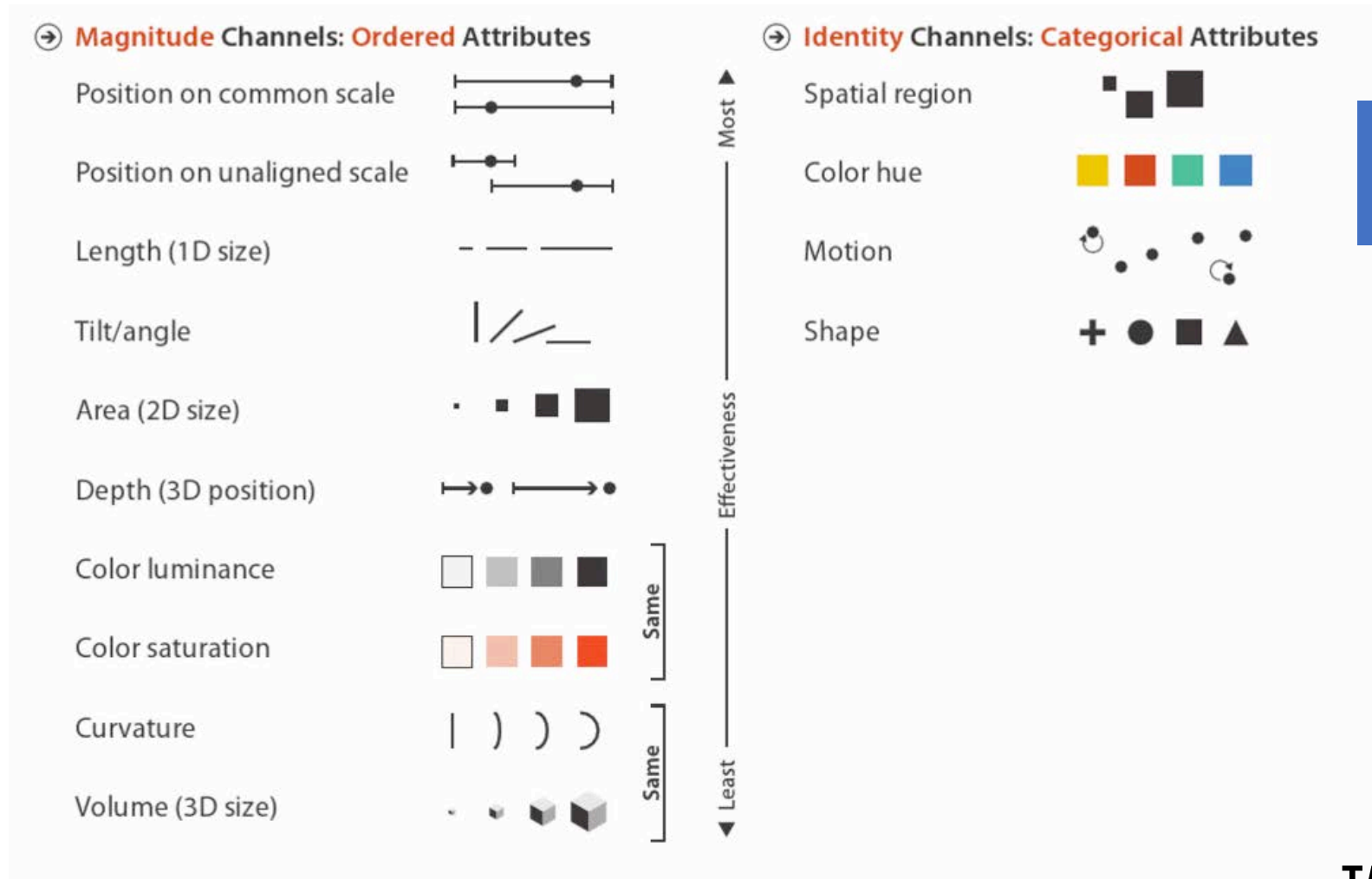
Visual Variable	Selective	Associative	Quantitative	Order	Length
Position	Yes	Yes	Yes	Yes	Dependant on resolution
Size	Yes	Yes	Approximate	Yes	Association: 5; Distinction: 20
Shape	With Effort	With Effort	No	No	Infinite
Value	Yes	Yes	No	Yes	Association: 7; Distinction: 10
Hue	Yes	Yes	No	No	Association: 7; Distinction: 10
Orientation	Yes	Yes	No	No	4
Grain	Yes	Yes	No	No	5
Texture	Yes	Yes	No	No	Infinite
Motion	Yes	Yes	No	Yes	Unknown

SO... WHAT SHOULD YOU USE?

If you want to distinguish **categories**, use...
position, color*, or maybe **shape**

If you want to show numbers use...
position or **length**, then **area**, then **color***

A FEW MORE INTERESTING QUIRKS



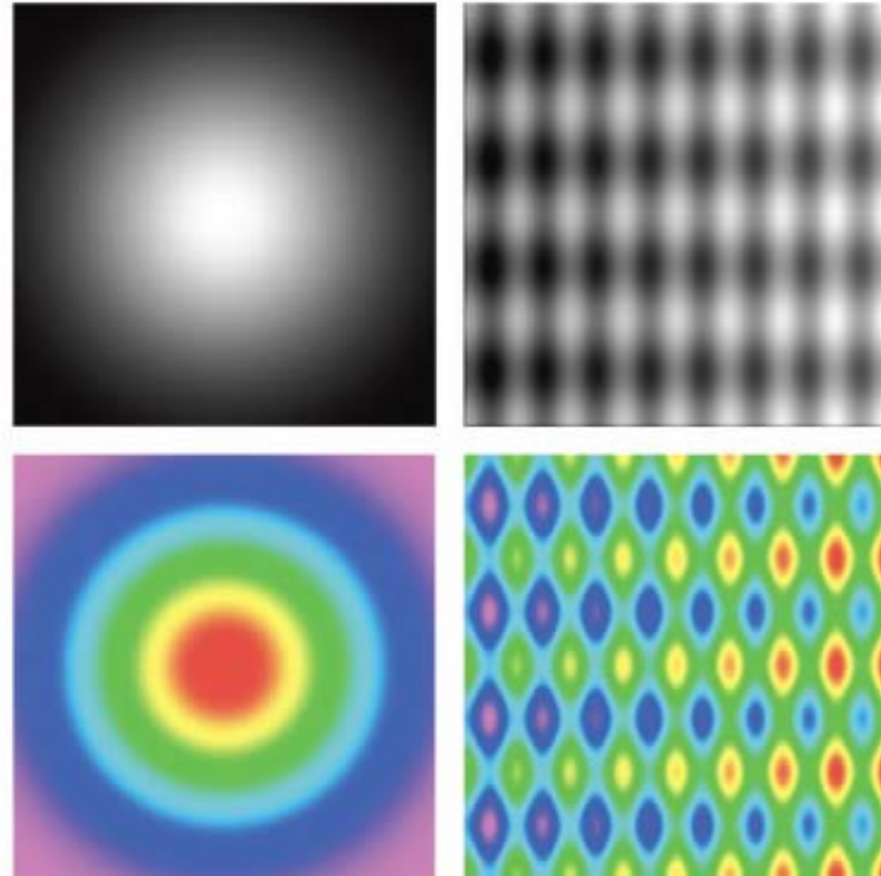
HUE

HUE IS GOOD FOR CATEGORIES ...AND BAD FOR EVERYTHING ELSE

CATEGORIES - GOOD*!



QUANTITIES - BAD!



*with a few caveats

A FEW MORE INTERESTING QUIRKS

AREA

BRIGHTNESS

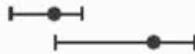
VOLUME

➔ Magnitude Channels: Ordered Attributes

Position on common scale



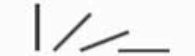
Position on unaligned scale



Length (1D size)



Tilt/angle



Area (2D size)



Depth (3D position)



Color luminance



Color saturation



Curvature



Volume (3D size)



➔ Identity Channels: Categorical Attributes

Spatial region



Color hue



Motion



Shape



Most

Effectiveness

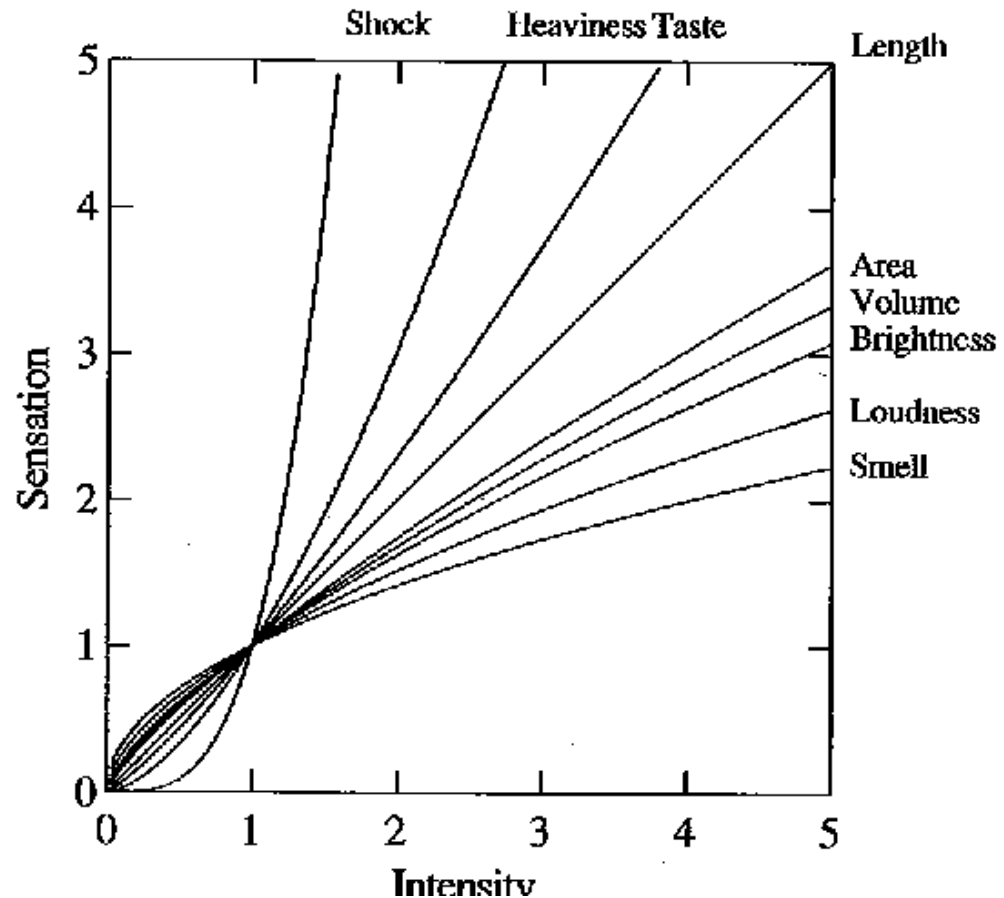
Least

HUE

SOME ENCODINGS ARE BAD BECAUSE WE DON'T PERCEIVE THEM LINEARLY

Relationship between stimulus and perception isn't always linear!

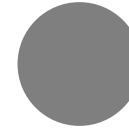
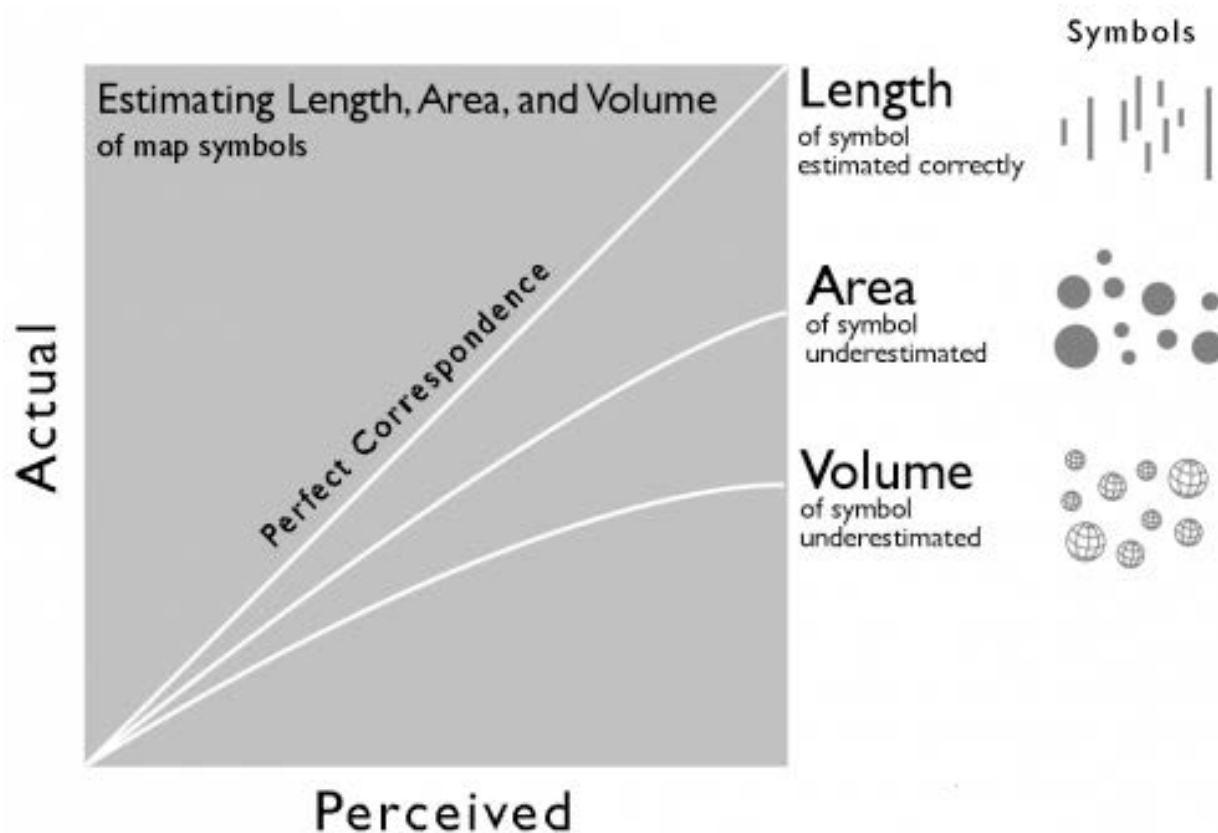
Stevens' power law describes a relationship between a physical stimulus (S) and its perceived intensity or strength (P).



NON-LINEAR PERCEPTION

People tend to **correctly** estimate lengths

They tend to **underestimate** areas and volumes.



When asked to pick a circle **2 times** the size, people tend to pick a circle **~1.8 times** larger.

This tendency **gets worse** as area grows.

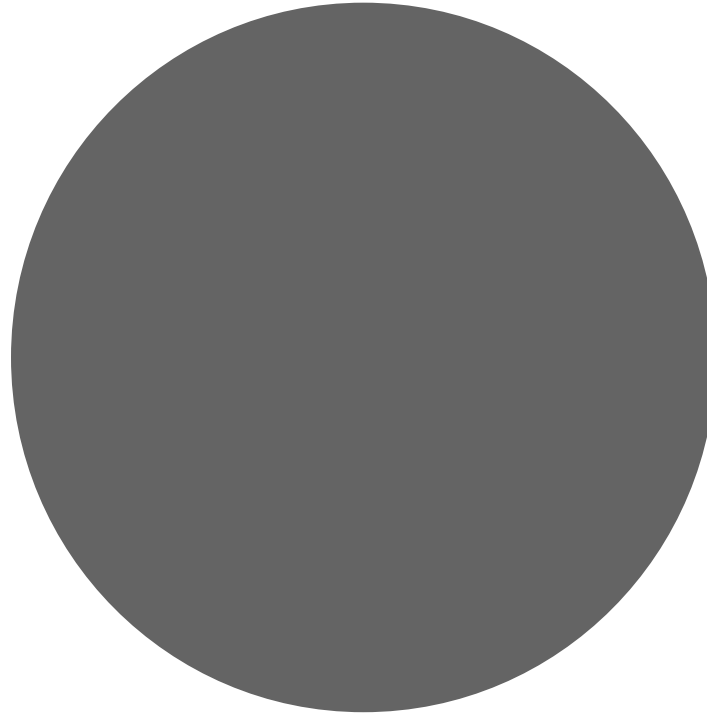
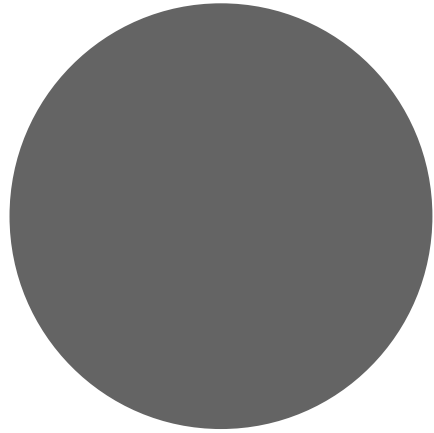
Volume is even worse!

TO ILLUSTRATE COMPARE THESE VALUES



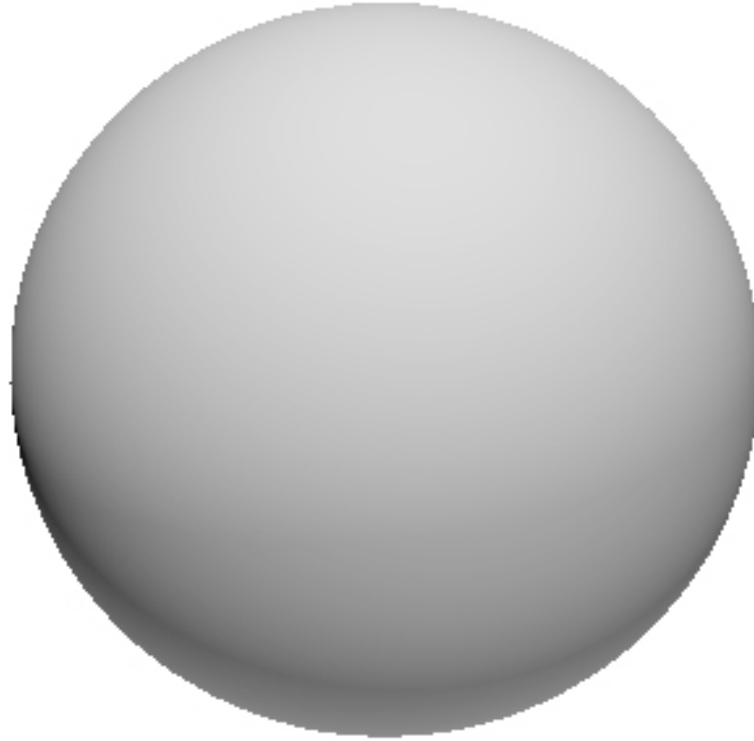
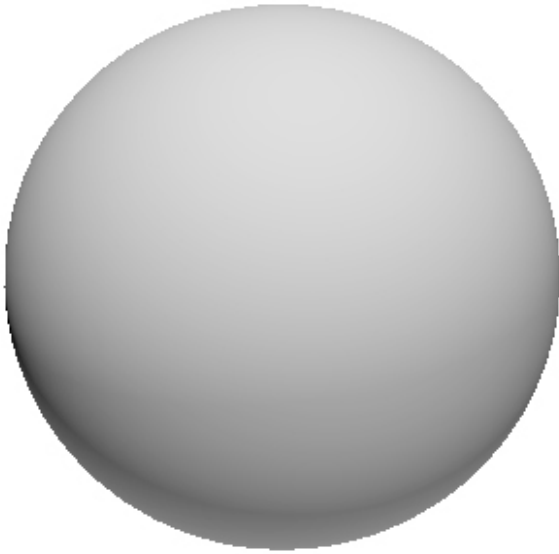
25% (1 : 4)

COMPARE CIRCLE AREAS



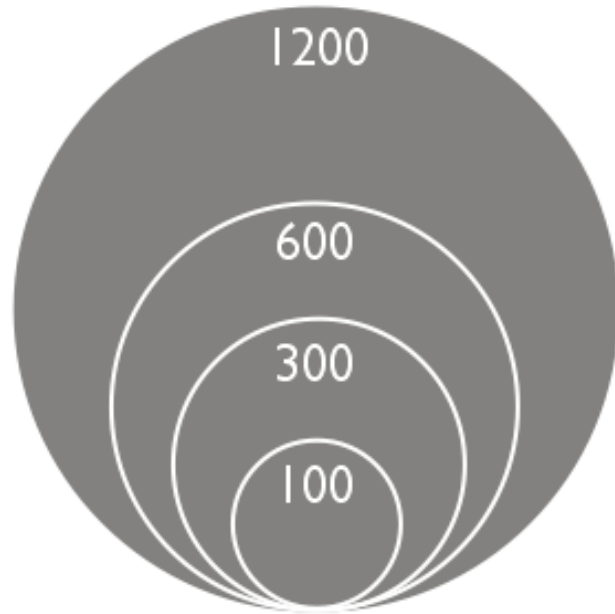
36% (1 : 2.8)

COMPARE SPHERE **VOLUMES**

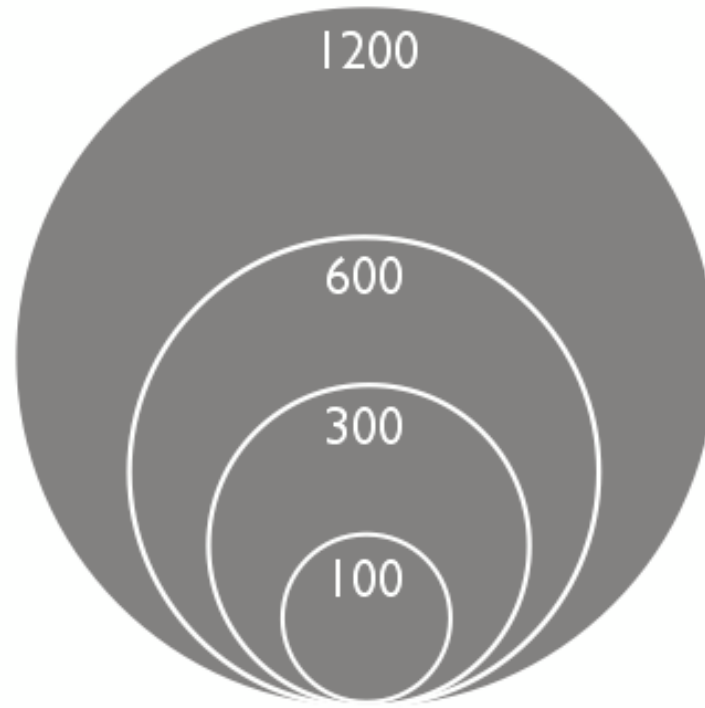


40% (1 : 2.5)

COMPENSATING FOR NON-LINEARITY



Absolute Scaling



Apparent Scaling
(Flannery's Compensation)

If we always underestimate the area, make everything bigger.

A FEW MORE INTERESTING QUIRKS

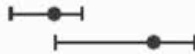
COMMON SCALES & BASELINES

➔ Magnitude Channels: Ordered Attributes

Position on common scale



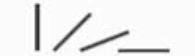
Position on unaligned scale



Length (1D size)



Tilt/angle



Area (2D size)



Depth (3D position)



Color luminance



Color saturation



Curvature



Volume (3D size)



➔ Identity Channels: Categorical Attributes

Spatial region



Color hue



Motion



Shape



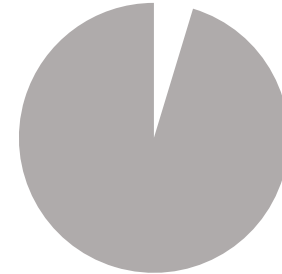
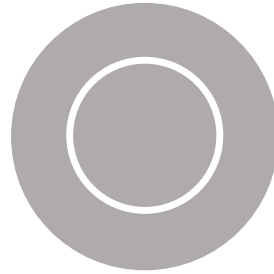
Most

Effectiveness

Least

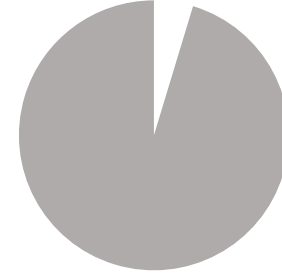
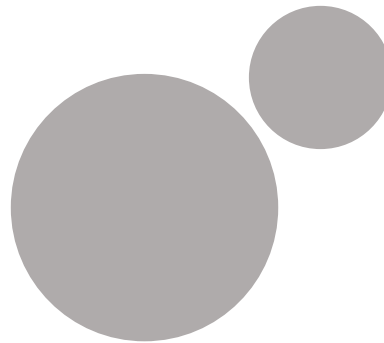
SCALES / BASELINES

**ALIGNED
(EASIER)**



**BUBBLE CHARTS AND PIE CHARTS ARE ALMOST
ALWAYS LESS USEFUL THAN BARS**

**UNALIGNED
(HARDER)**

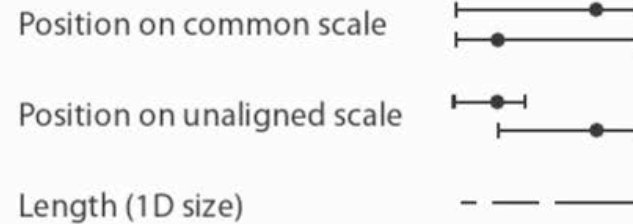


SO DON'T USE THEM

(Unless you're prepared to give a strong, specific, and perceptually-grounded rationale.)

FOR THE PURPOSES OF THIS COURSE...

➔ Magnitude Channels: Ordered Attributes



➔ Identity Channels: Categorical Attributes



Unless you're designing a new visualization, you rarely need to think about any encodings other than these.

Most
Effectiveness
Least