

Scientific and Large Data Visualization

19 October 2018

Information Visualization Basics

Massimiliano Corsini

Visual Computing Lab, ISTI - CNR - Italy

InfoVis – Next Lessons

- **Information Visualization – Introduction and Motivations**
- **Data Types, Graph Types and Visual Perception**
- **Visualization in Python**
- **Multidimensional Data**
- **Time Series**
- **Graph Drawing**
- **Practice**
 - **D3.js**

InfoVis – Basics

- **Introduction and motivations**
- **Ingredients of effective visualization**
- **Data types**
- **Graph types**
- **Visual perception**

Information Visualization

- Information visualization is the study of (interactive) visual representations of abstract data to reinforce human cognition. [Wikipedia]
- The use of computer-supported, interactive, visual representations of abstract data to amplify cognition. [Card et al. 1999]
- The use of computer graphics and interaction to assist humans in solving problems. [Purchase et al. 2008]

Information Visualization

- The purpose of information visualization is to amplify cognitive performance, not just to create interesting pictures. [Card 2007]
- *Infographics* is a visual tool for communication, for the understanding and for the analysis. [Alberto Cairo]

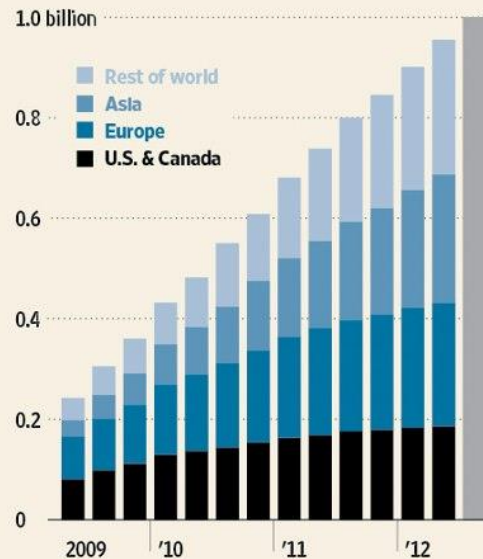
InfoGraphics – Charts

Facebook Nation | The social network reaches 1 billion active users

If Facebook were a nation, it would have the third-largest population. But if you compared its revenue to national economies, it would rank 156th.

Total population, billions of people		Ranking		2011 gross domestic product, billions of dollars
1.34	China	#1	#154	Swaziland 3.98
1.24	India	#2	#155	Fiji 3.81
1 billion users*	Facebook	#3	#156	Facebook \$3.71 billion
0.31	U.S.	#4	#157	Barbados 3.69
0.24	Indonesia	#5	#158	Togo 3.59

Facebook's monthly active users by region, quarterly (current breakdown not available)



Note: For purposes of reporting MAUs and revenue by geographic region, Europe includes Russia and Turkey, Asia includes Australia and New Zealand, and rest of world includes Africa, Latin America, and the Middle East.

InfoGraphics – Diagrams

A LAS PUERTAS DEL CIELO

“La iglesia del Sagrado Corazón de Jesús es una metáfora que se refiere a la existencia humana y su relación con Dios. Esto queda en evidencia a simple vista: a medida que la obra alcanza altura, los personajes (representados por estatuas) están más cerca del cielo”. Así la explica Carmelo D’Agostino, el más antiguo de los capuchinos de Córdoba.

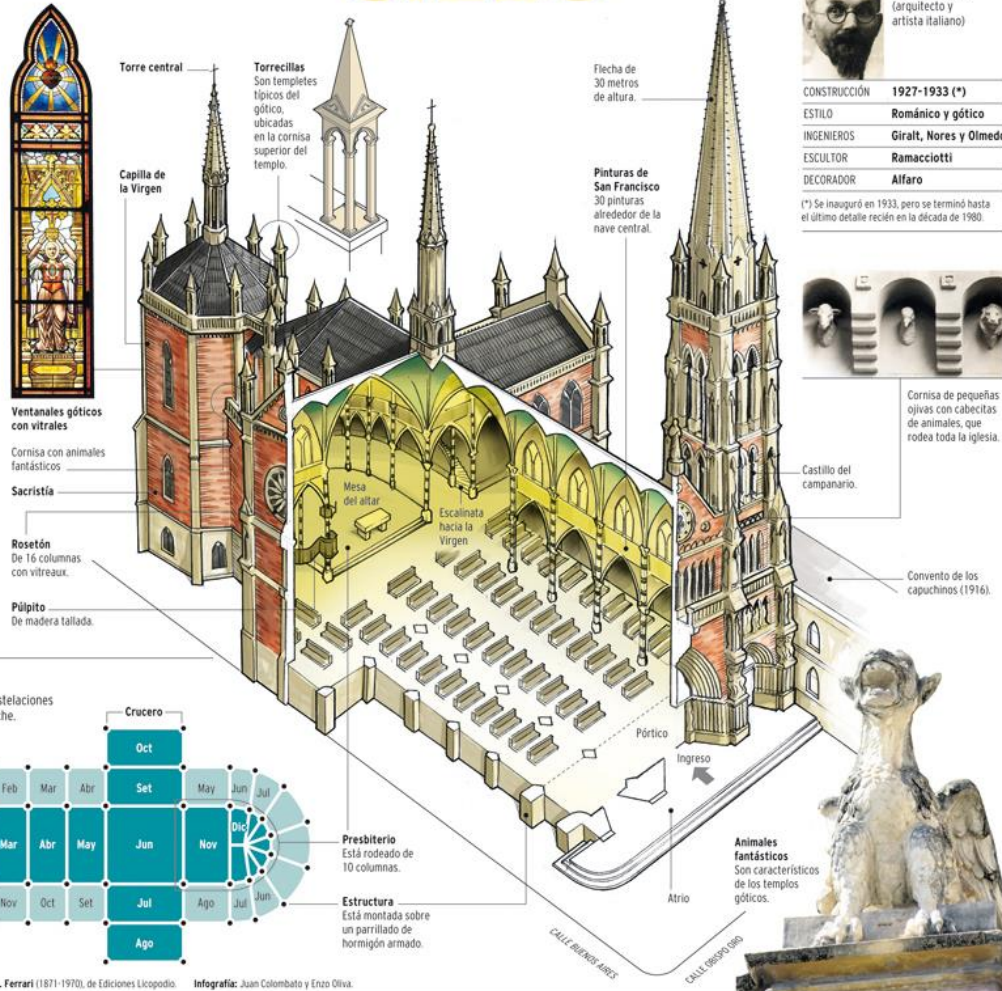
En la parte más baja, al nivel del zócalo, se ven arañas, tortugas, escorpiones y sapos, que simbolizan el pecado. Sobre ellos se yerguen columnas con diferentes formas, símbolos de las diversas culturas anteriores a Cristo. En un nivel superior se encuentran los atlantes, que son los hombres que cargan con las torres de la iglesia. Inmediatamente arriba de los atlantes, el arquitecto Augusto Ferrari ubicó a los doce apóstoles, representantes de las virtudes que Dios espera alcancen los seres humanos. Luego están los santos, que lograron esas virtudes y viven en la gloria de Dios. A la misma altura de los santos, está la estatua del Sagrado Corazón. Por último, en lo más alto, en medio de las torres, Ferrari colocó a San Francisco, para que desde ahí custodie a toda la ciudad.

BOVEDA ESTRELLADA

En el cielo raso, están pintadas las constelaciones del firmamento de Córdoba a medianoche.



Fuente: párroco Sebastián Glasman - Augusto C. Ferrari (1871-1970), de Ediciones Licopodio. Infografía: Juan Colombata y Enzo Oliva.



FICHA TÉCNICA

	AUTOR Augusto Ferrari (arquitecto y artista italiano)
CONSTRUCCIÓN	1927-1933 (*)
ESTILO	Románico y gótico
INGENIEROS	Giralt, Nores y Olmedo
ESULTOR	Ramacciotti
DECORADOR	Alfaro

(*) Se inauguró en 1933, pero se terminó hasta el último detalle recién en la década de 1980.



Cornisa de pequeñas ojivas con cabezitas de animales, que rodea toda la iglesia.



ORNAMENTOS DE LA FACHADA

Combina elementos góticos y románicos. Está construida de manera que, a medida que alcanza altura, los personajes se encuentran más cerca del cielo.



Estatua de San Francisco
Desde ahí custodia toda la ciudad.

REFERENCIAS

- Estatuas
- Parejas de atlantes
- Templetes
- Columnas multiformes
- Columnas lisas

SÍMBOLOS HACIA EL INFINITO

ESPIRITUS CELESTES CRIADOS POR DIOS PARA SU MINISTERIO

Estatuas de ángeles

LOS QUE LOGRARON LAS VIRTUDES Y VIVEN EN LA GLORIA DE DIOS

Cornisa de cabezas de animales

Estatua de la caridad

REPRESENTANTES DE LAS VIRTUDES

Apóstoles

SÍMBOLOS DEL PECADO

Atlantes

Arañas, tortugas, escorpiones y sapos al pie de las columnas.

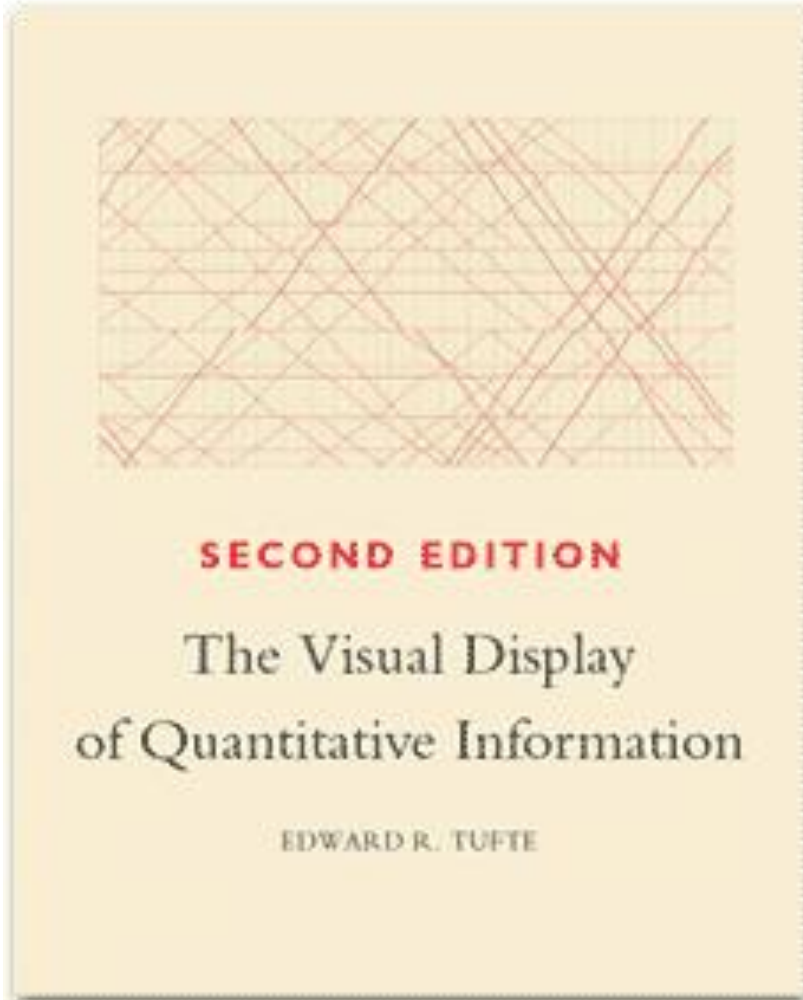
Columnas multiformes
A los costados de los portales, representan las razas de todos los pueblos.

Columnas lisas
Son para dar la impresión de fuerza.

Information Visualization

- Difference from *Scientific Visualization*:
 - Information visualization treats also abstract data (numerical and non-numerical data).
 - In scientific visualization spatial representation is given.
- Difference from *Visual Analytics*:
 - In Visual Analytics the accent is on the reasoning/interaction loop.

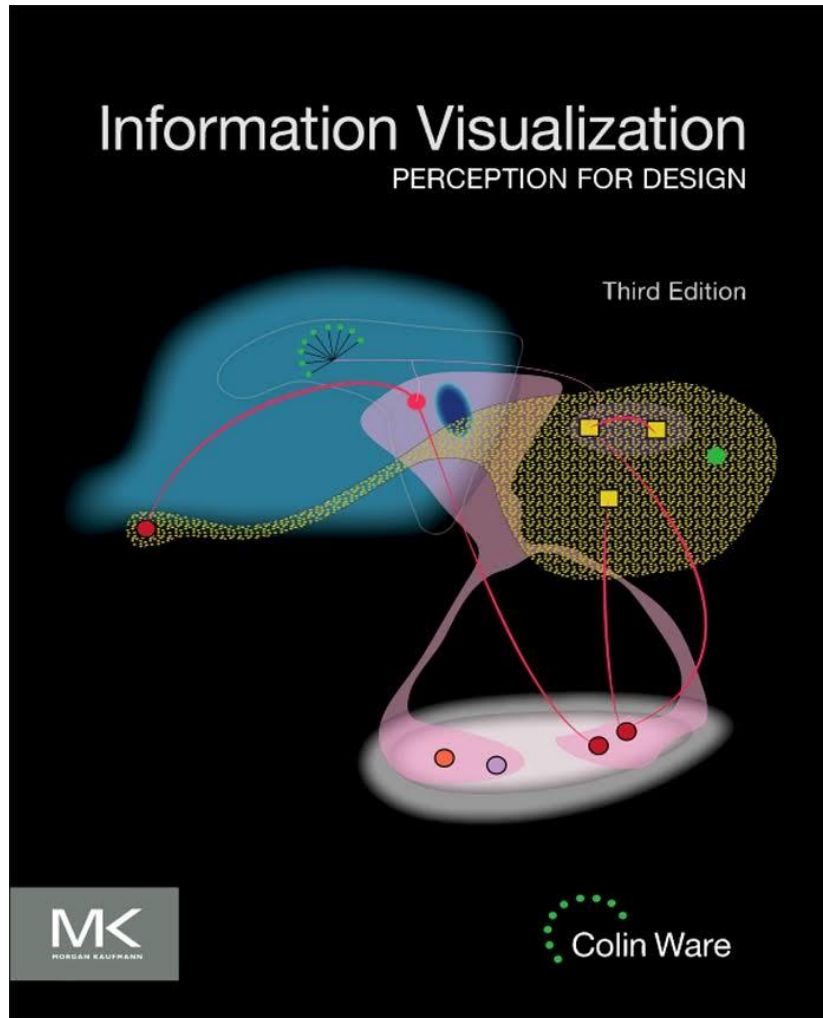
Data Visualization



A classic book on statistical graphics, charts, tables. Theory and practice in the design of graphics to visualize data.

Very practical, very classic.

Information Visualization

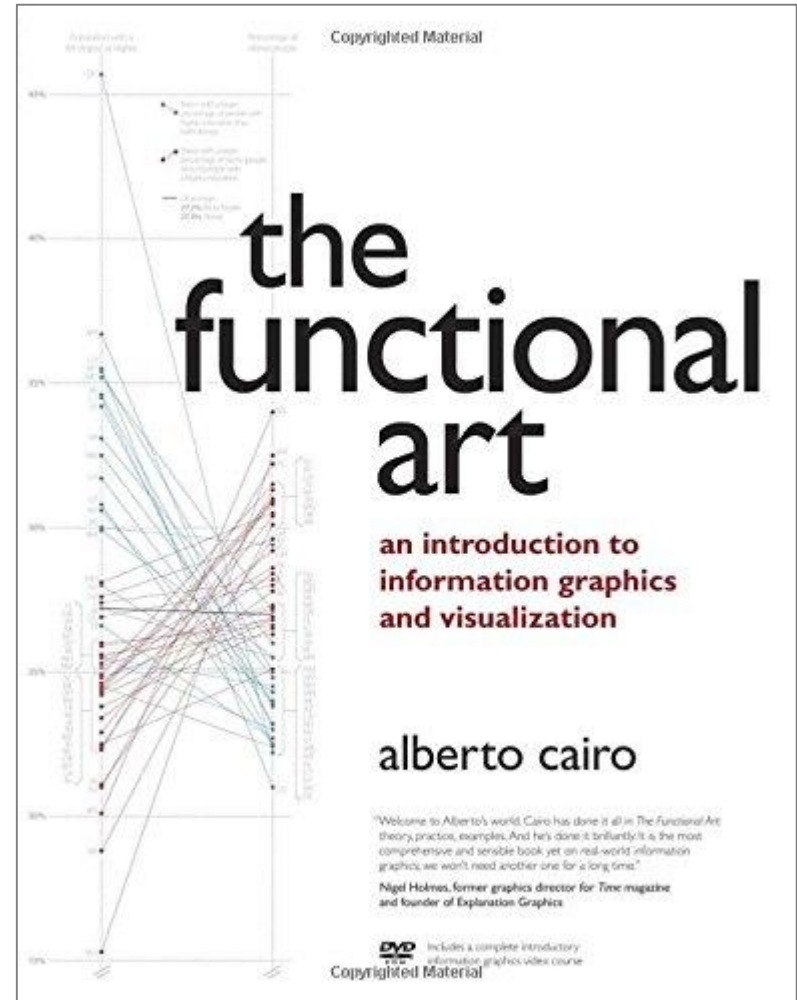


A very interesting aspect:

why we need to
consider visualization
as a *SCIENCE*

Functional Art

- Function does not *dictate* but *restrict* our choices.
- This is particularly true for infographics.



Motivations

Unemployment rate (%)

	CURRENT	Historical maximum	Historical minimum
Alabama	6.7	14.4	3.3
Alaska	7.5	11.5	5.9
Arizona	6.9	11.5	3.6
Arkansas	6.2	10.2	4.1
California	9.3	11.0	4.7
Colorado	6.1	9.1	2.5
Connecticut	7.1	10.0	2.1
Delaware	6.1	8.2	2.9
Florida	8.1	9.7	3.3

Motivations

Unemployment rate (%)

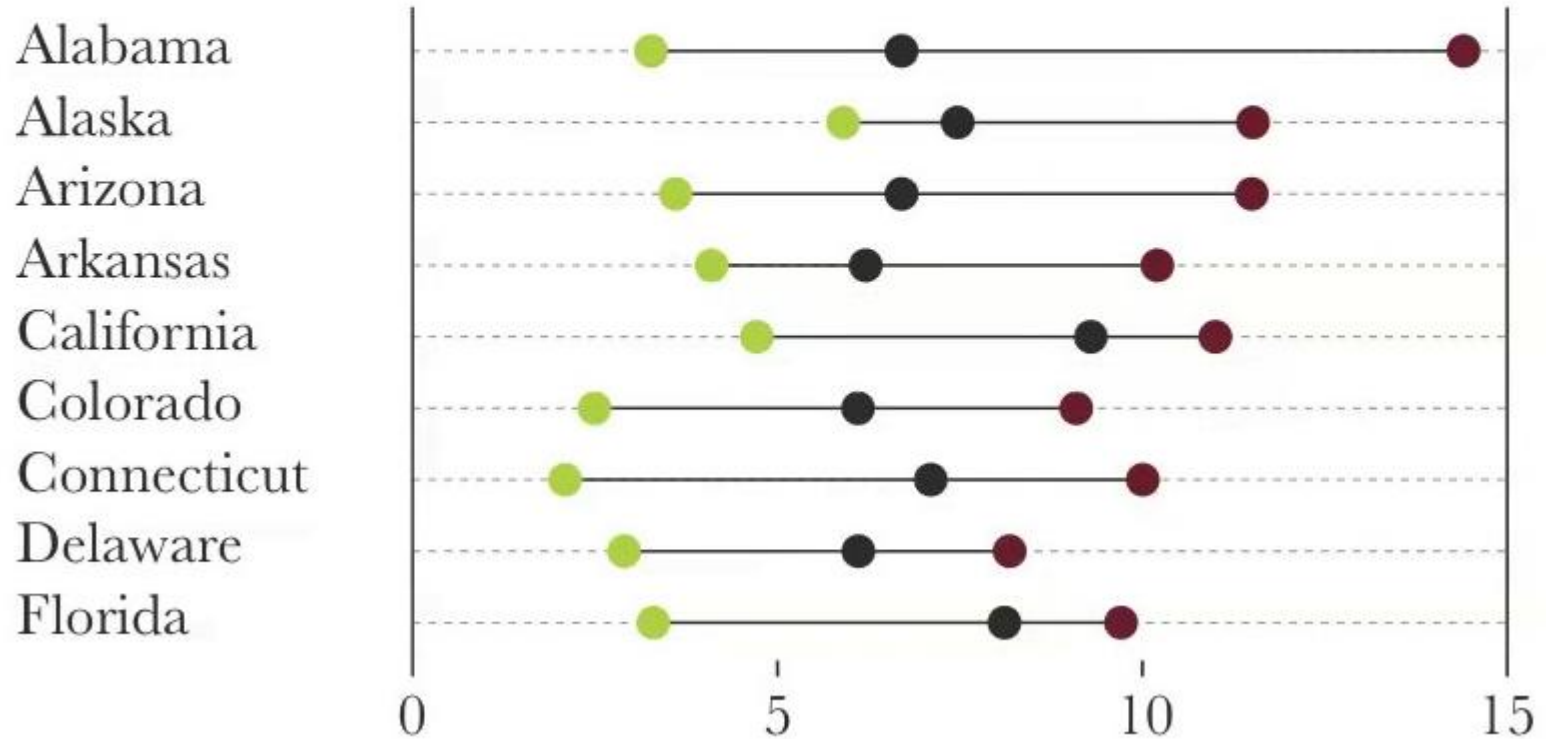
	CURRENT	Historical maximum	Historical minimum
Alabama	6.7	14.4	3.3
Alaska	7.5	11.5	5.9
Arizona	6.9	11.5	3.6
Arkansas	6.2	10.2	4.1
California	9.3	11.0	4.7
Colorado	6.1	9.1	2.5
Connecticut	7.1	10.0	2.1
Delaware	6.1	8.2	2.9
Florida	8.1	9.7	3.3

Which country is close to its historical maximum ?

Motivations

Unemployment rate (%)

● Current ● Historical maximum ● Historical minimum



Easier to answer... Why ?

Rationale

- The human visual system (HVS) is very good at identifies and analyzes patterns.
- We can visualize data to make easy for our brain to analyze them, for example to do comparisons.

Effectiveness

- To be effective, data visualization should be taken into account several factors:
 - Data type
 - Goal (function)
 - Visual Perception System

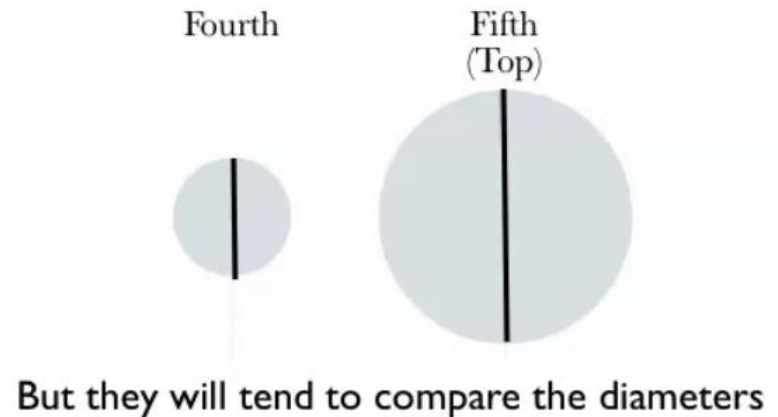
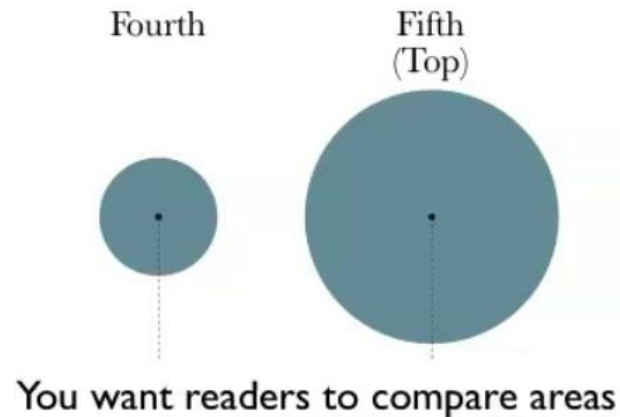
Example – Comparisons

How much you would save if federal income taxes were reduced 20%



SOURCE: The New York Times

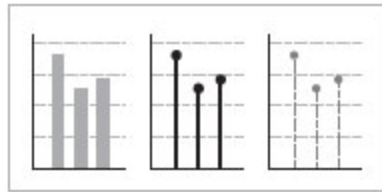
GRAPHIC: ACME



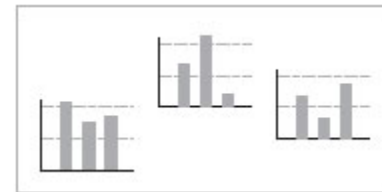
Cleveland and McGill 1984

Less accurate →

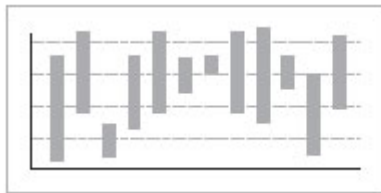
Less accurate ↓



Position along a common scale



Position along nonaligned scales



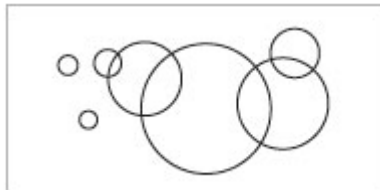
Length



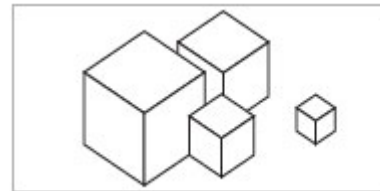
Direction



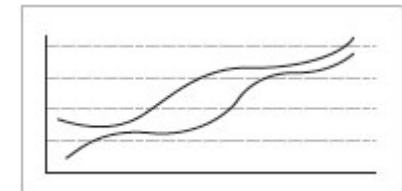
Angle



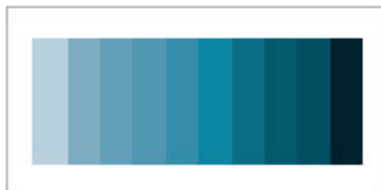
Area



Volume



Curvature



Shading



Color saturation

Adapted from *"The functional art"* by Alberto Cairo

Key Ingredients

- In the following we focus on:
 - Data types
 - One dimension, two dimensions, N-dimensions
 - Quantitative, Ordinal, Nominal
 - Graph types
 - Visual Perception
- We give some design guidelines time by time.

Data

- Information is obtained from data (!)
- Structured / Unstructured.
- Generated by sensors, by computers, by humans, etc.

Variable Types

- According to Stevens (1946):
 - Nominal
 - Labels (e.g. apples, oranges, bananas)
 - Ordinal
 - To ordering things (e.g. ranks of movies)
 - Interval
 - Interval scale of measurements (e.g. time of departure-arrival)
 - Ratio
 - Measures defined on a ratio scale (e.g. the mass of an object)

S. S. Stevens, “*On the theory of scales and measurements.*”, *Science*, 103, pp. 677-680, 1946.

Variable Types

- Category
 - Steven's nominal class (e.g. country names, type of disease)
- Ordinal
 - Labels expressing degree (e.g. cold, hot, very hot)
 - In general, encoded as integer data.
- Quantitative
 - Intervals, measures, etc.
 - In general, real-numbered data.

Data Dimensions

- Common dimensions:
 - Univariate, bivariate, trivariate
 - Multi-variate ($N > 3$ dimensions)
- Variables can be dependent or independent.
- Each case is a point in a space with N dimension (*data point*).

Data Dimensions

- A set of data can be represented by a table with N columns (one for each variable).

	Variable 1	Variable 2	Variable 3
Data 1			
Data 2			
Data 3			
Data 4			
...			

Data Relationships

- A table can be used to represent a relationship between different data.

	User Id	Game id
Data 1	Smith	023923
Data 2	James	238548
Data 3	Frank	385753
Data 4	Powell	357352
...		

Data Relationships

- 1-to-1
- 1-to-many
- Many-to-many
- Relationships may also have attributes

What we want..

- A set of well formed and interconnected tables of data.

Table 1		

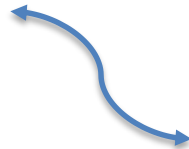


Table 2		

Table 3		



What we have..

- Data does not come in the form we would like.
- Data may have inconsistencies:
 - Corrupted data
 - Missing data
 - Several data may be equivalent (e.g. text field “R&D”, “Research”, “Research and Development”, “r*d”)

Data Processing Pipeline

- Data does not come in the form we would like.
- Typical data processing pipeline (from raw data to clean structured data):
 1. Collect data.
 2. Data simplification (extract a subset of interest).
 3. Clean and structure them.
- An output of the pipeline can be also *metadata*.

Data Collection

- Search and download data.
- Parse text.
- Convert between different formats.
 - Examples: CVS to JSON , MySQL to HTML , etc.
- Merge heterogeneous sources.

Data Simplification

- Filter / Selection
 - Remove unwanted data
 - Remove invalid data (null values)
- Aggregation
 - Collapse several data points into a single one
 - Replace some values with minimum, maximum, average, total, etc.

Data Processing Pipeline

- Typically performed automatically using *scripts*.
- Human-guided data transformations is possible (through macro-operations)
 - Use appropriate tools (e.g. OpenRefine - <http://openrefine.org>)

Metadata

- Data which describes the data
 - Role of variables
 - Type of variables
 - Constraints
 - Dependencies
- Collection and processing operations can be also described.

How to Present Data ?

- How to present data graphically ?
 - To allow visual analysis
 - To highlight patterns
 - To answer some specific questions
 - Etc.

Quantitative Values

- We have just mentioned the work by Cleveland & McGill (1984)
 - Position
 - Length
 - Angle / slope
 - Area
 - Volume
 - Color saturation / shading



**Perceptual
Accuracy**

QUANTITATIVE	ORDINAL	NOMINAL
Position	Position	Position
Length	Density	Hue
Angle	Saturation	Texture
Slope	Hue	Connection
Area	Texture	Containment
Volume	Connection	Density
Density	Containment	Saturation
Saturation	Length	Shape
Hue	Angle	Length
	Slope	Angle
	Area	Slope
	Volume	Area
		Volume

From “Data Visualization” Course by John C. Hart, for Coursera, 2015.

Table vs Graphs

- Present tables directly is preferred when:
 - *Few* data points
 - Precise values are important

Univariate Data

- Few interesting solutions.
- Statistical description:
 - Mean, median, standard deviation, quartiles.
- Warning! (some data that appears to be univariate are actually bivariate).

Stemplots

- Also called *stem and leaf* plot.
- Used to display quantitative data, generally from small data sets (50 or fewer observations).
- Easy to print.
- Easy to read.

STEMS	LEAVES
0	5 8
1	2 3 5 7
2	0 0 0 5 8 8 9
3	0 0 1 3 3 3 6 6 7 7 7 7 7 8 8 8 8 9 9
4	1 3 5 5 5 6 7 7 8 8 8 8 9 9
5	0 0 0 1 1 1 1 2 6 8
6	0 0 1 1 2 4 4 4 4 4 8 8 9
7	0 5 5 5 5 7
8	3 4 4 5 6 6 6 7 8 9
9	0 1 2 2 2 2 5 5 6 8 9 9
10	2 2 2 5 7

Raw Data in Row:

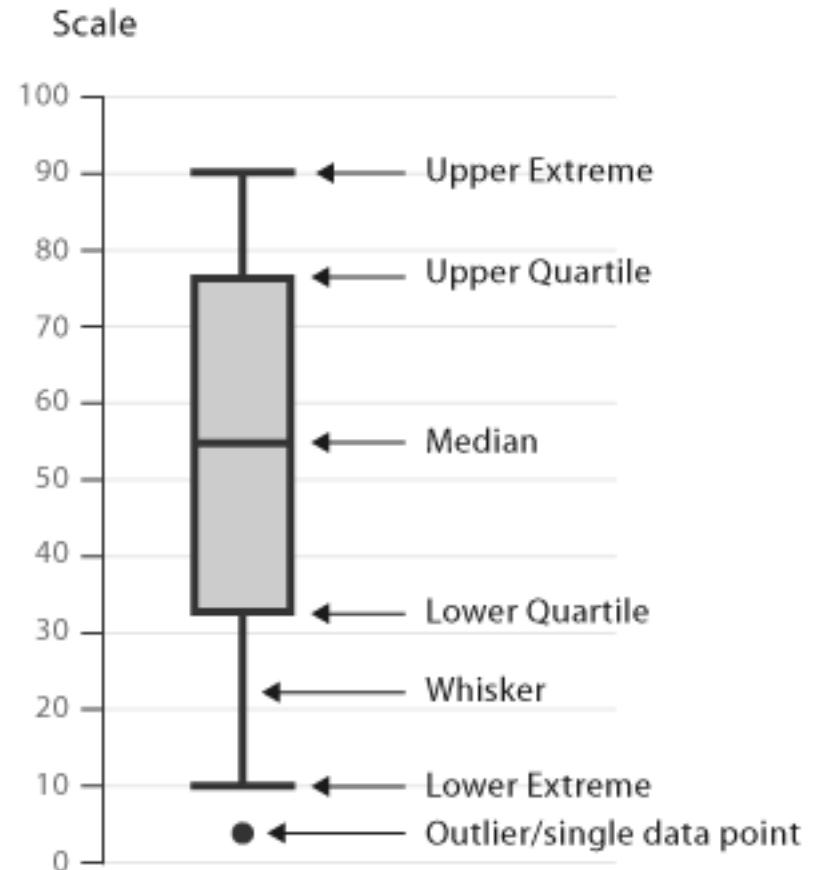
102, 102, 102, 105, 107

Stemplots

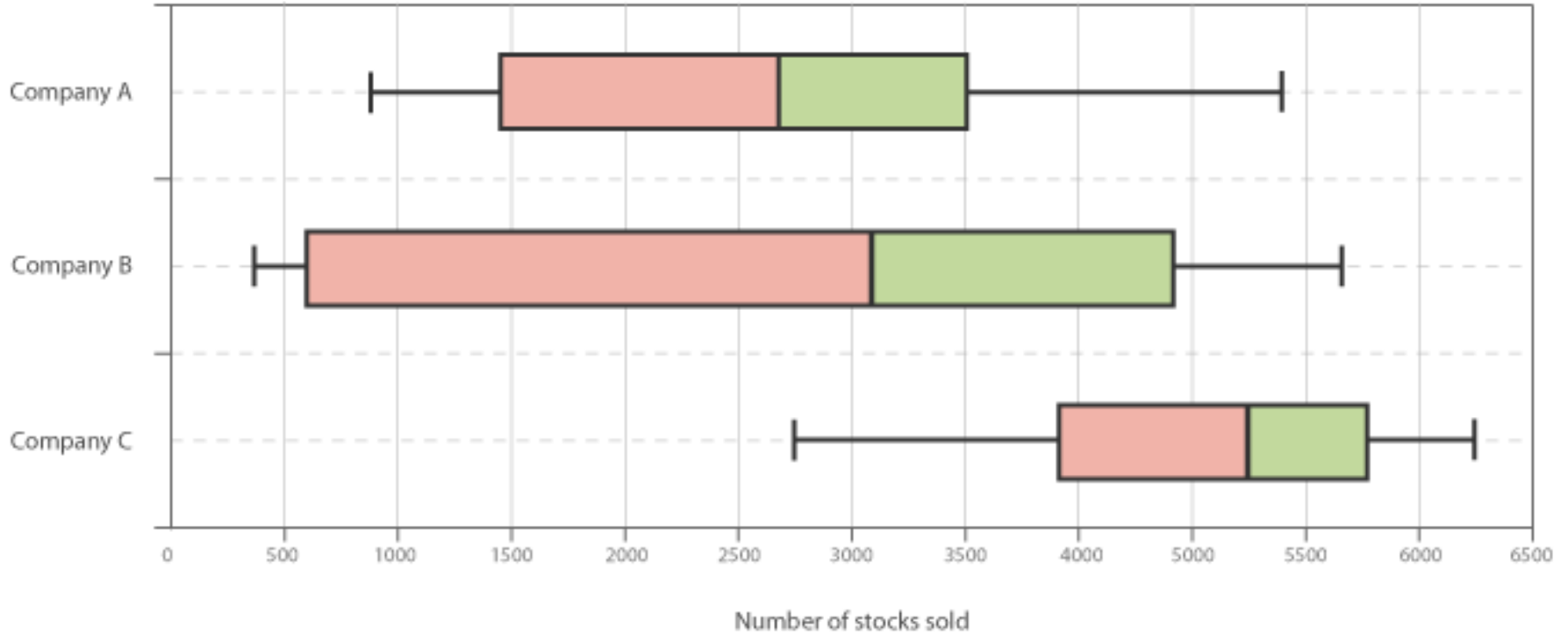
Northbound	Hour	Southbound
45 20 03	5	05 48
55 49 32 20 13 01	6	02 23 35 57
58 53 49 44 38 32 25 19 13 08 02	7	00 07 16 20 26 30 37 46 52 59
59 57 54 50 47 44 39 35 31 28 24 21 18 14 09 05 00	8	01 08 12 17 21 29 31 35 39 44 49 53 58
52 48 44 39 34 29 23 18 12 05	9	03 10 18 27 32 37 45 51 58
53 47 41 37 32 27 22 15 07	10	00 07 14 21 30 39 48 57
55 49 35 29 23 16 08 01	11	06 11 19 27 34 41 50 59
56 48 44 39 32 27 21 14 05	12	02 15 30 45 57
50 45 35 30 25 20 15 05	13	03 10 18 23 29 37 45 56
52 43 32 24 12 03	14	00 09 18 27 39 48 57
58 44 31 26 15 06	15	01 17 29 41 55
56 40 30 22 11	16	10 25 38 50
55 41 32 23 14 01	17	00 20 34 53
58 49 42 36 28 22 16 09	18	05 14 21 29 37 45 56
57 51 46 39 33 28 23 17 13 08 02	19	02 09 14 19 23 27 32 36 40 44 48 53 57
52 43 30 21 15 06	20	09 17 26 34 40 49 55
45 30 16 03	21	10 20 30 40 50
50 30 10	22	15 35 55

Box and Whisker Plots

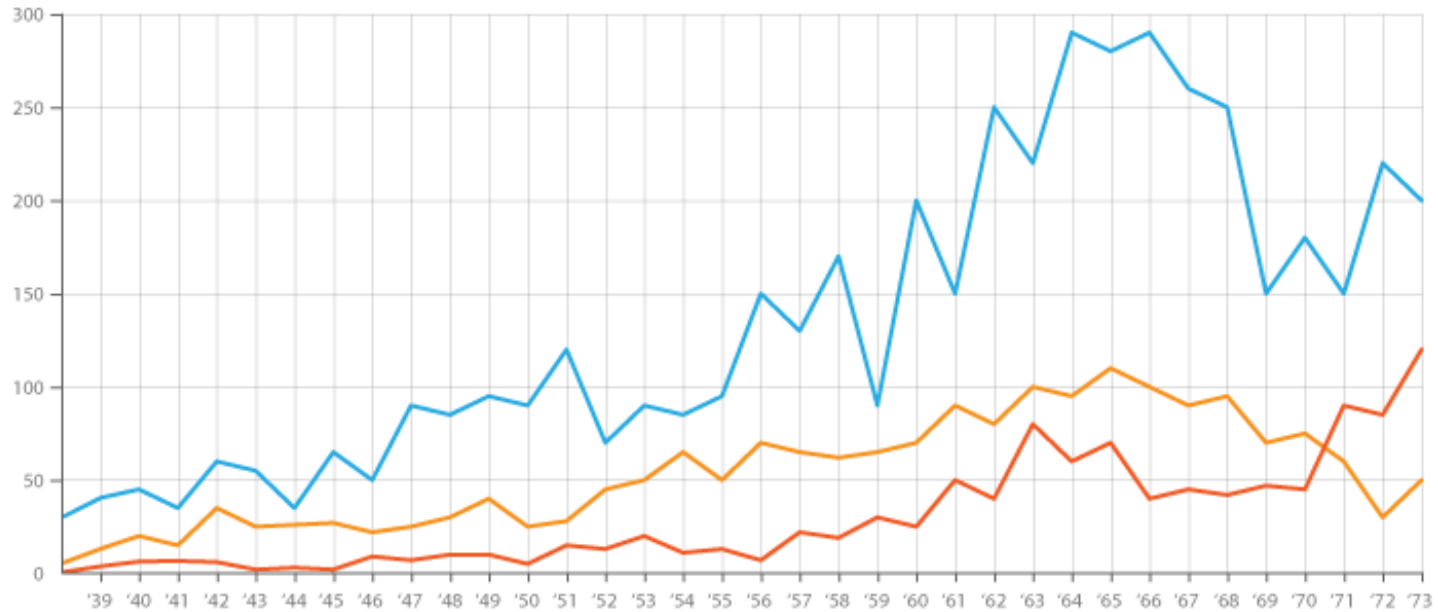
- *Box and Whisker Plot (or Box Plot)* is a convenient way of visually displaying a data distribution through their quartiles.
- **Advantages:**
 - Key values (average, median, 25th percentile, etc.)
 - If there are any outliers and what the values are.
 - If the data is skewed and in what direction.



Box and Whisker Plots



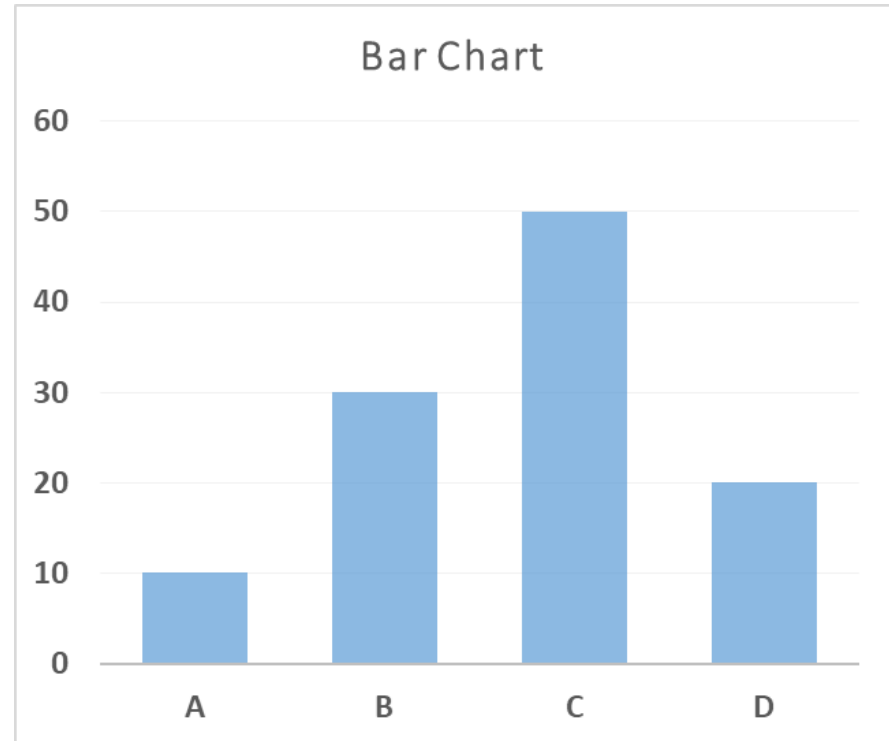
Line Charts/Line Graphs



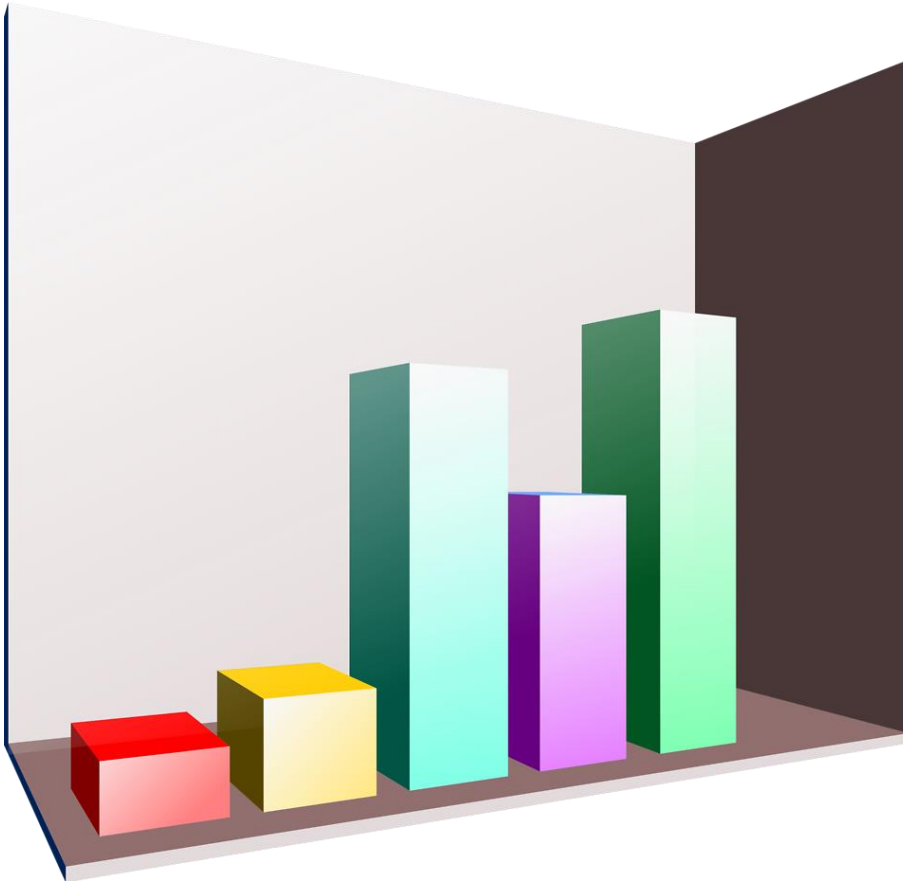
- Invented by the Scottish engineer and statistician William Playfair (1759-1823)
- Two quantitative variables, typically:
 - $X \rightarrow$ time or intervals , Y any
- Line indicates that there are also intermediate values

Bar Charts

- Bivariate Data
 - One nominal variable (typically independent)
 - One quantitative variable (typically dependent variable)
- Horizontal/Vertical bars
- Do not confuse with *histograms*.



Bar Charts

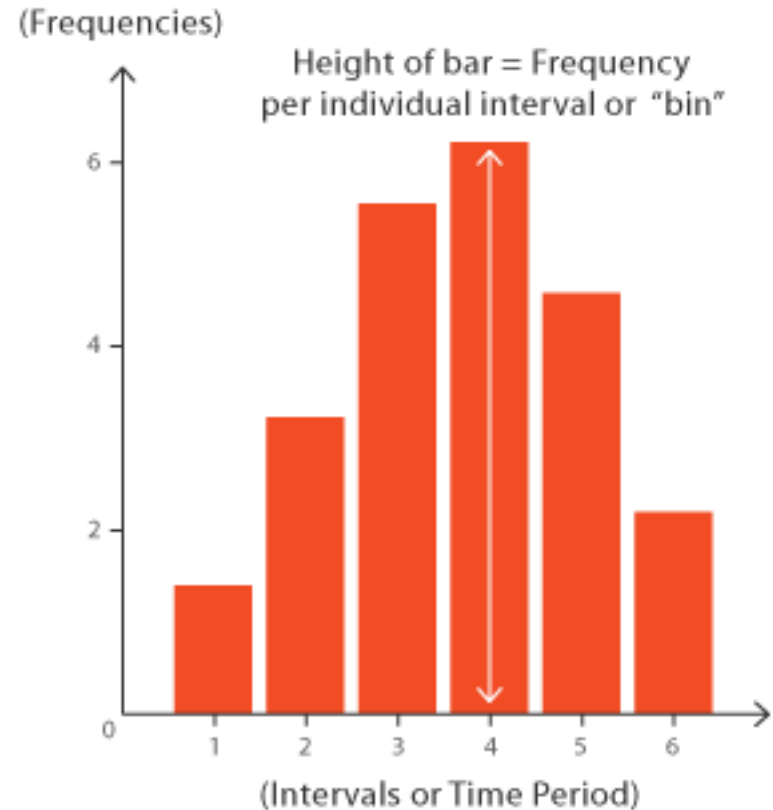


3D?!

Not a very good idea..

Histograms

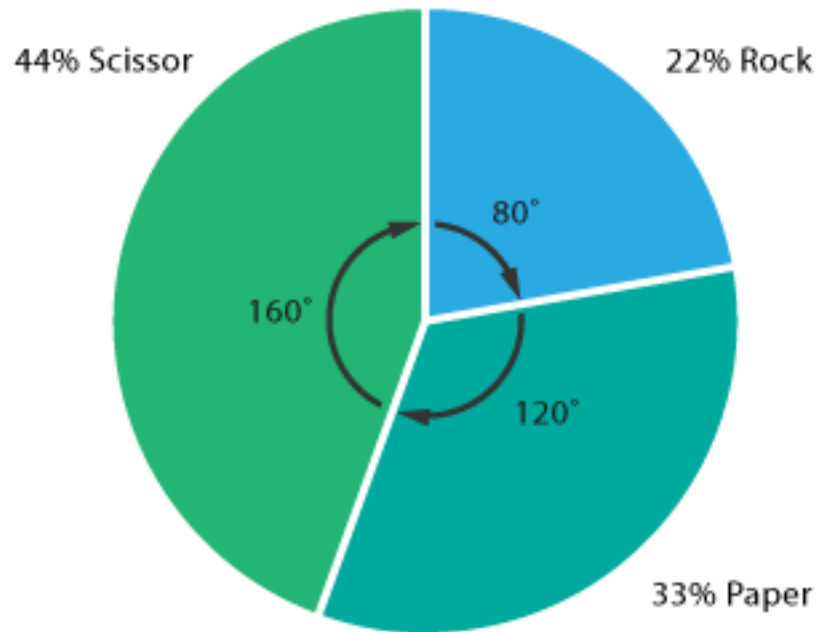
- Bivariate Data
 - One independent and one dependent variable
 - The first variable is quantized in intervals (bins)



Pie Charts

- Bivariate Data
 - One independent and one dependent variable
- Good for a quick visual check.
- Not good for:
 - Many values.
 - Accurate comparisons.

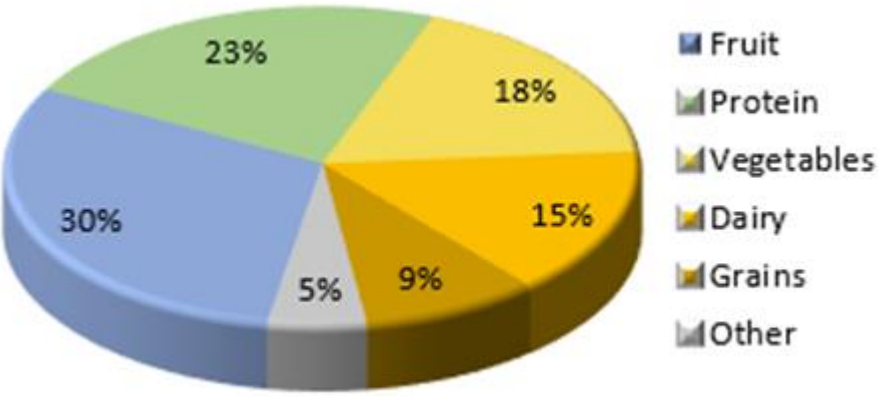




Data			
Rock	Paper	Scissor	TOTAL
2	3	4	9
To calculate percentages			
$2/9=22\%$	$3/9=33\%$	$4/9=44\%$	100%
Degrees for each "pie slice"			
$(2/9) \times 360$ = 80°	$(3/9) \times 360$ = 120°	$(4/9) \times 360$ = 160°	360°

Figure from *Data Visualization Catalogue* (<http://datavizcatalogue.com>)

Pie Charts



**3D?!
Not more readable.**

Quando o brasileiro come fora

O crescimento da economia muda os hábitos alimentares e estimula o mercado de comida pronta

Luiz Salomão e Alberto Cairo

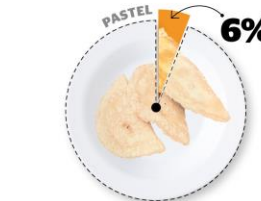
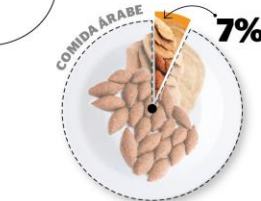
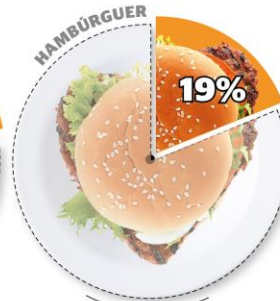
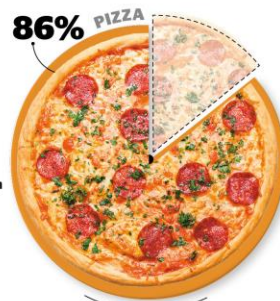
O **BRASILEIRO** vai cada vez menos à cozinha - e nisso já se aproxima dos países desenvolvidos. Segundo o estudo *Alimentação fora do lar na visão do consumidor brasileiro*, da GS&MD, consultoria de consumo especializada em varejo, **mais de 30% do gasto dos brasileiros com alimentação é feito com serviços de entrega ou refeições fora de casa**. Só Estados Unidos, Portugal, Reino Unido e Espanha estão à frente. Esse percentual era de 24% em 2002. Oitenta e quatro por cento dos 1.224 entrevistados em grandes capitais disseram que costumam comprar alimentos prontos para consumo, seja em supermercados, padarias ou restaurantes. A pesquisa atribuiu essa mudança, em primeiro lugar, ao crescimento econômico: o Brasil cresceu três vezes mais que a média das economias mais desenvolvidas nos últimos cinco anos, e a classe C já representa quase metade da população. Em segundo lugar, as mulheres já são 43% do total de trabalhadores. Por fim, os "domicílios unipessoais" passaram de 3,4 milhões em 1996 para 6,3 milhões em 2006.

As campeãs de entregas
Porcentual de entrevistados que declararam pedir o alimento citado "com regularidade". Mais de 50% dos pesquisados dizem encomendar comida mais de uma vez por semana

Nas capitais

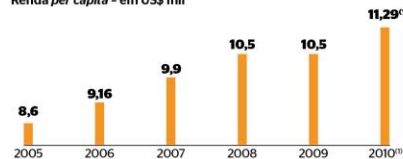


A pizza é campeã em todo o país, mas sobretudo (sem surpresas) em São Paulo. A comida chinesa faz sucesso no Recife



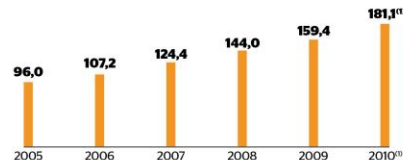
O aumento da renda no país...

Renda per capita - em US\$ mil



...acompanha o crescimento das empresas de comida pronta e restaurantes

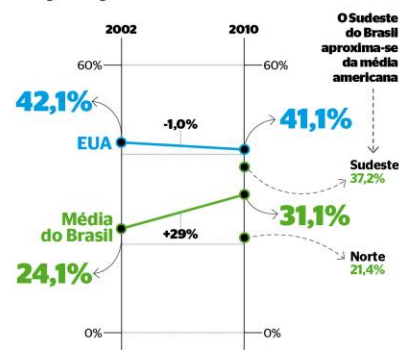
Faturamento anual total - em R\$ bilhões



(1) Estimativa. Fontes: FMI, Abia

Porcentual do orçamento familiar para alimentação gasto em refeições fora de casa

Média no Brasil e nos EUA e em algumas regiões brasileiras



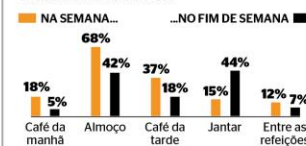
Quem gasta mais fora de casa

% sobre o total das despesas das famílias com alimentação



Em dias úteis, almoço fora; no fim de semana, jantar fora

Entrevistados que declaram comer habitualmente fora de casa



O brasileiro valoriza o sabor, a higiene e a aparência

Atributos que definem a qualidade da refeição (mais de uma resposta possível)



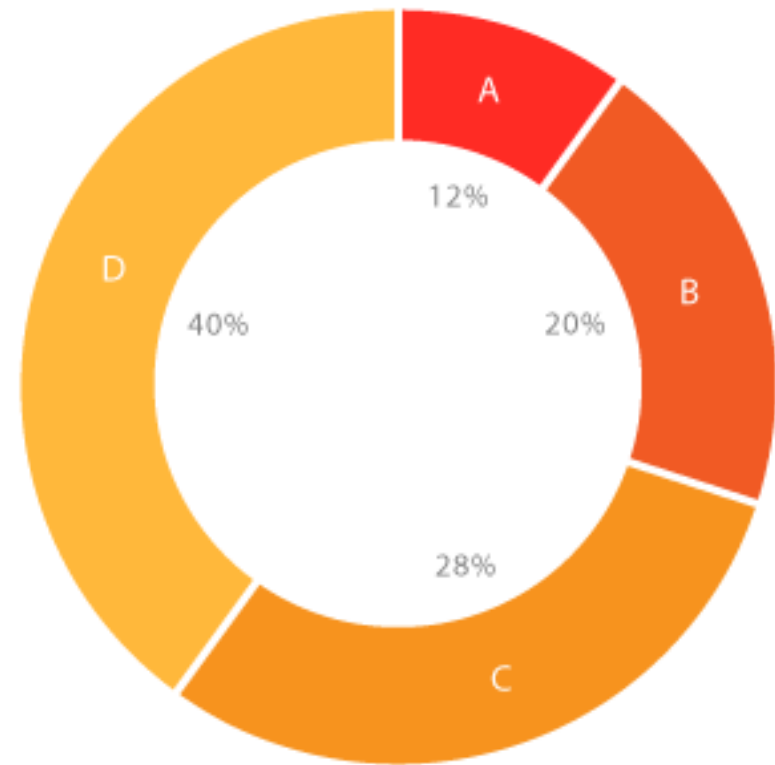
Fatores que determinam o gasto fora de casa

Por classe



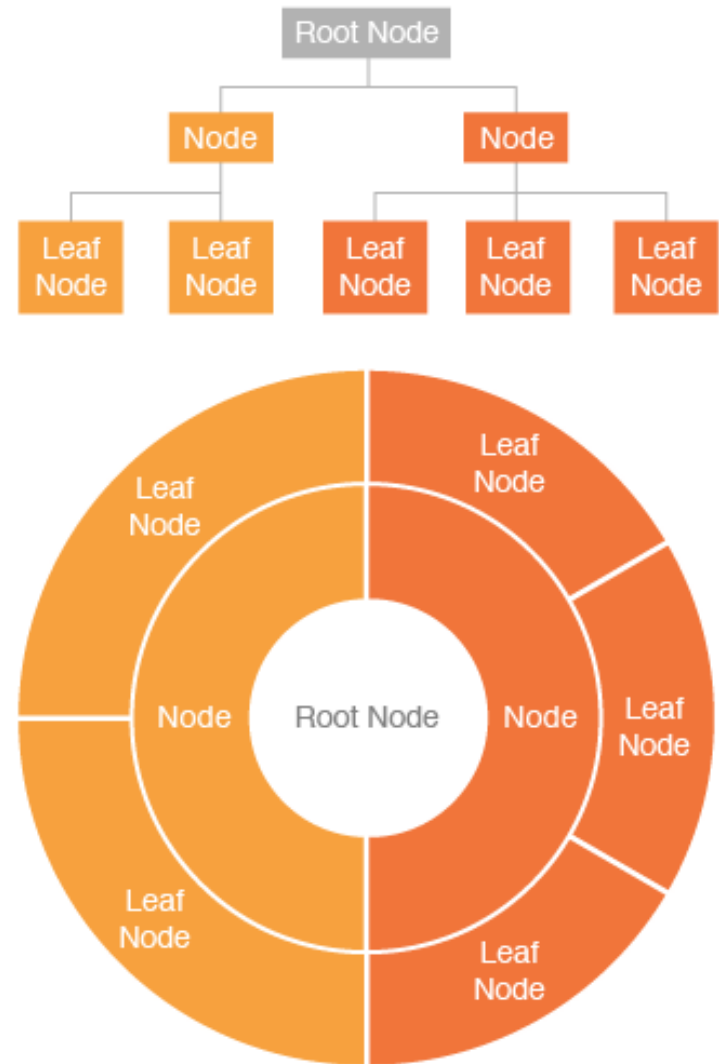
Donut Charts

- Essentially, a pie chart with the center area cut out.
- Allows to focus more on arc length instead of comparing the proportion between slices.
- More space efficient.

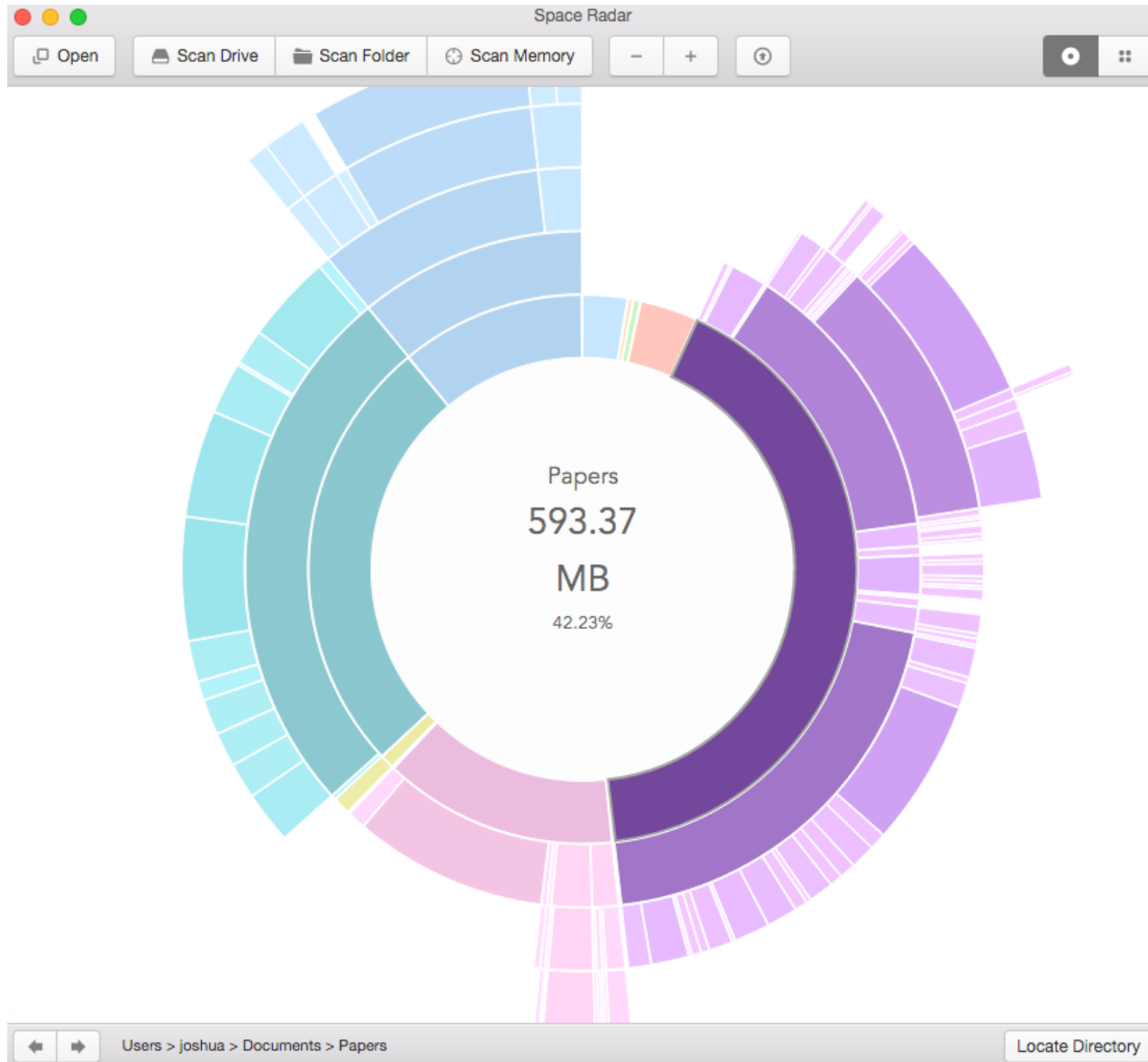


Sunburst Charts

- To show *hierarchy* through a series of rings. Each ring corresponds to a level in the hierarchy.
- Hierarchy moving outwards from the center.
- Colour can be used to highlight hierarchical groupings or specific categories.



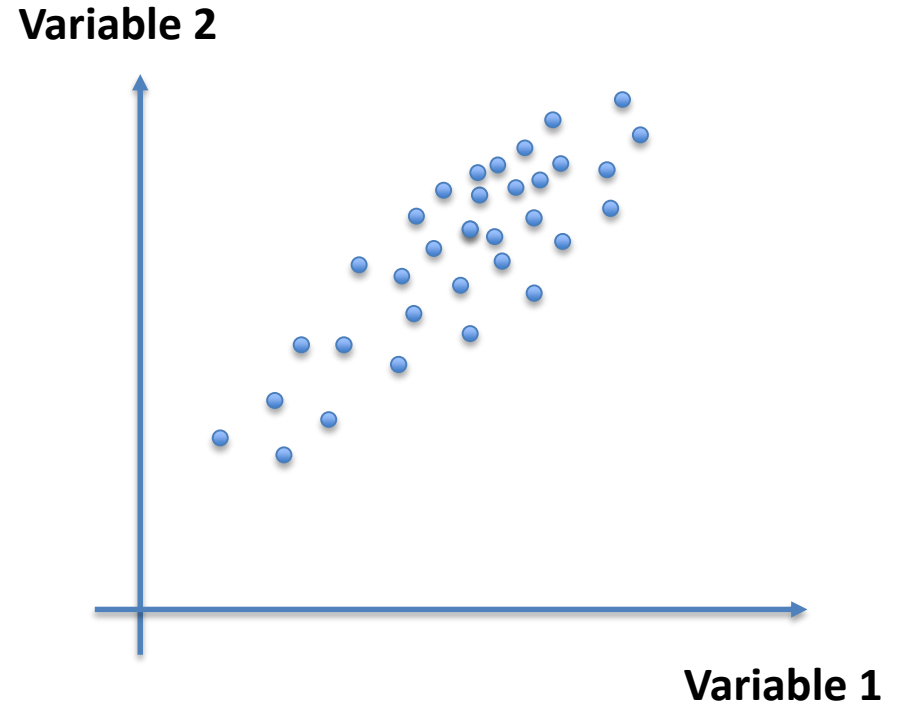
Sunburst Charts



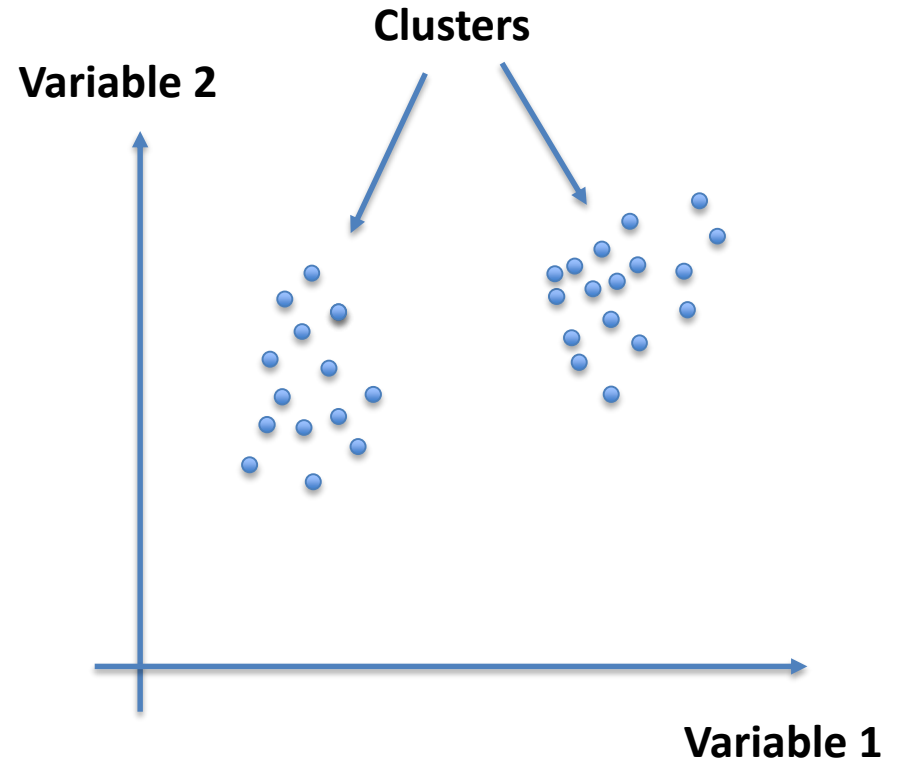
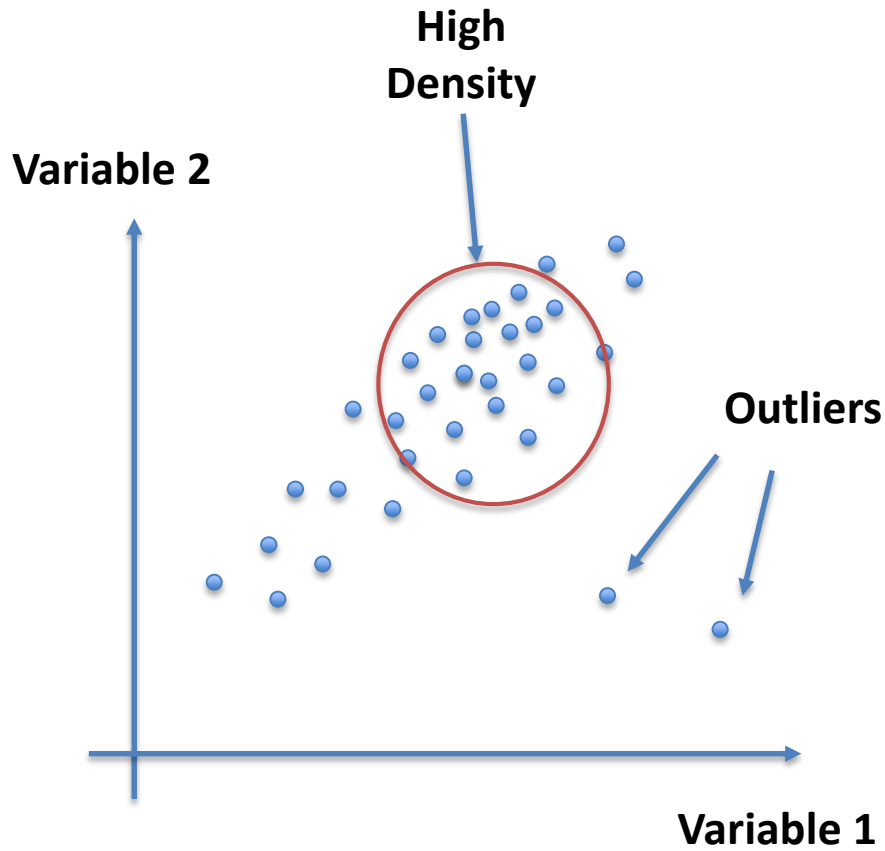
Produces by **Space Radar** app (<https://github.com/zz85/space-radar>)

Scatter Plots

- Bivariate Data
 - Two independent variables
- Good to identify relationships, outliers and clusters.

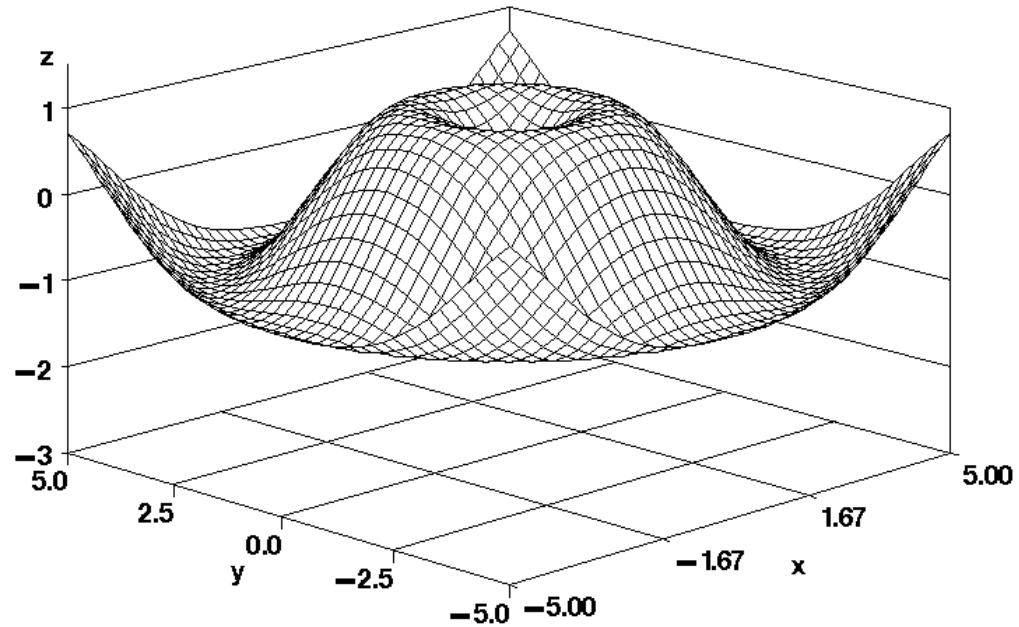


Scatter Plots



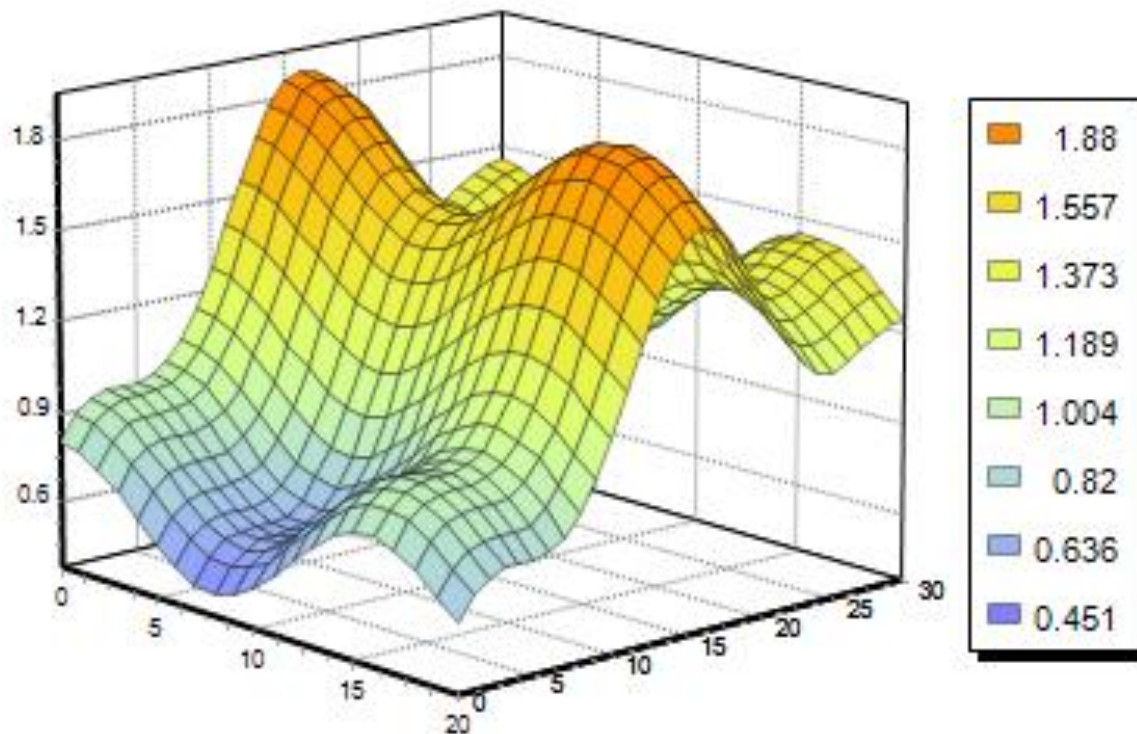
Surface Graphs

- Trivariate Data
 - Three continuous variables
 - Two independent and one dependent



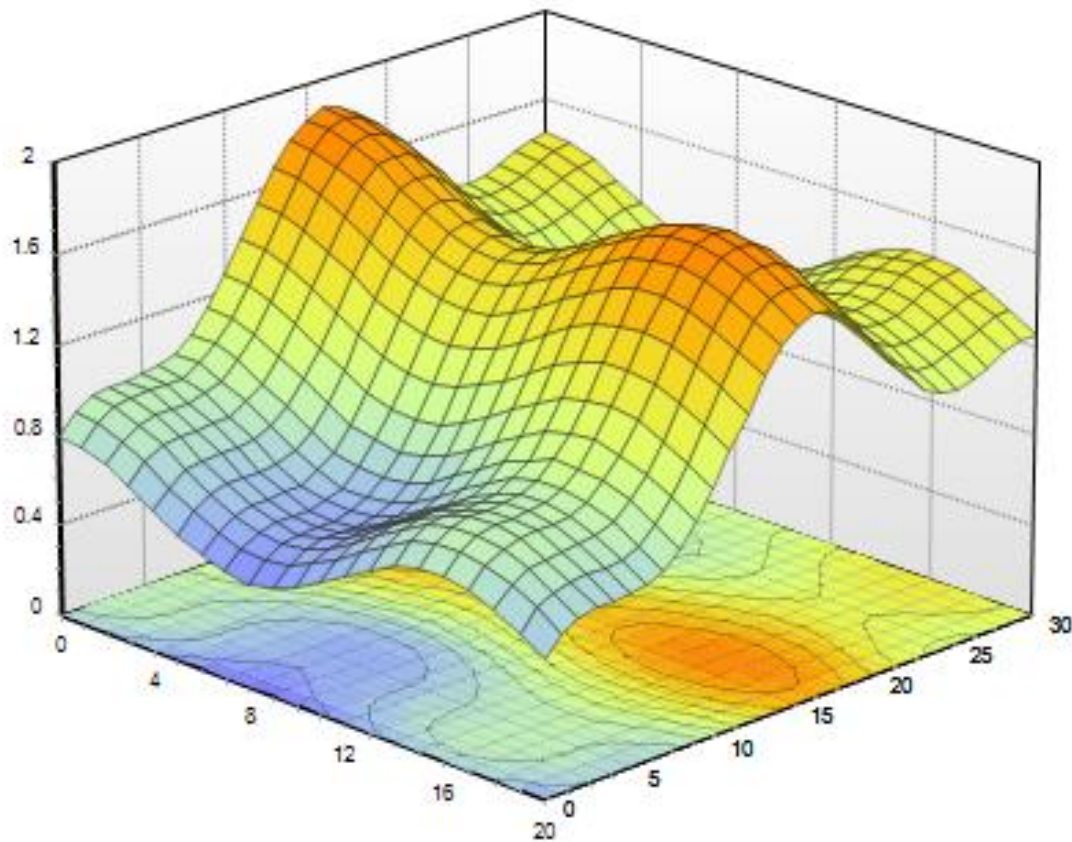
Surface Graphs

- Color may be associated to the dependent variable.



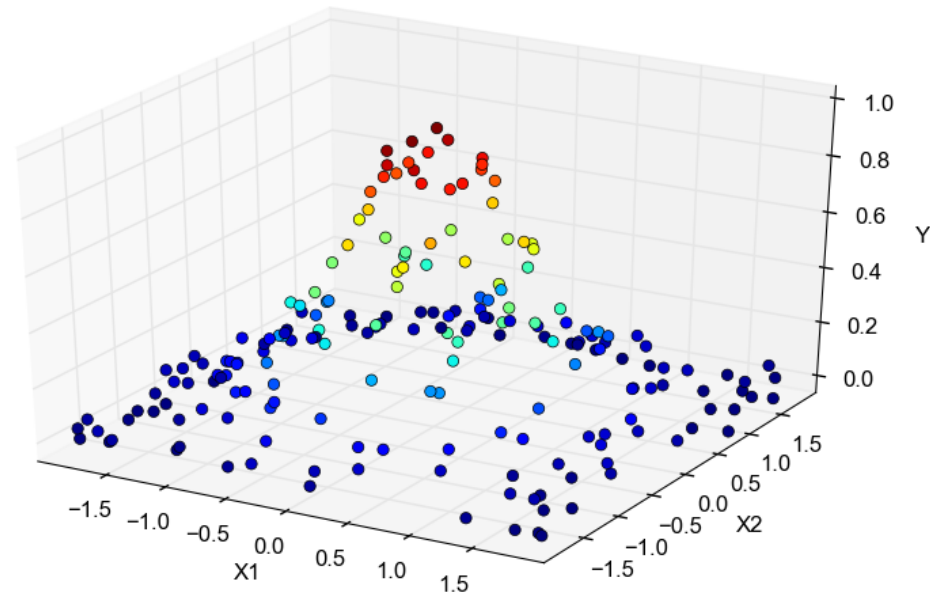
Surface Graphs

- Level curves can be also used.



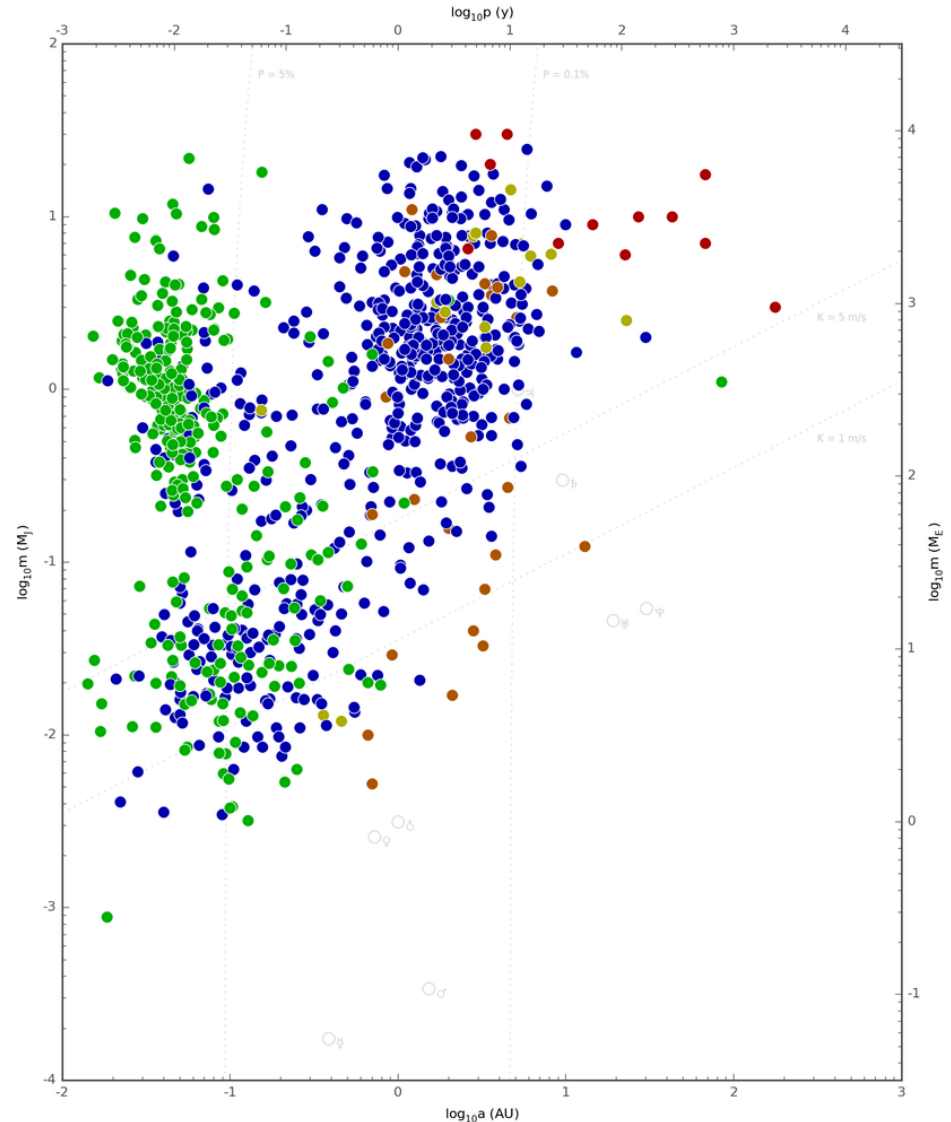
3D Scatter Plots

- Trivariate Data
 - Three quantitative variables
- Same concept of 2D Scatter Plot



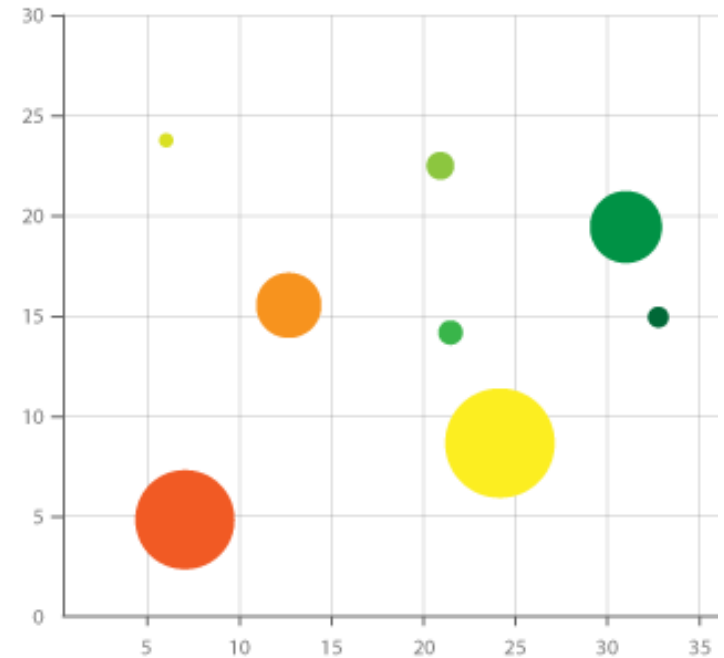
Colored Scatter Plots

- Trivariate Data
- Color can encode a variable or a category



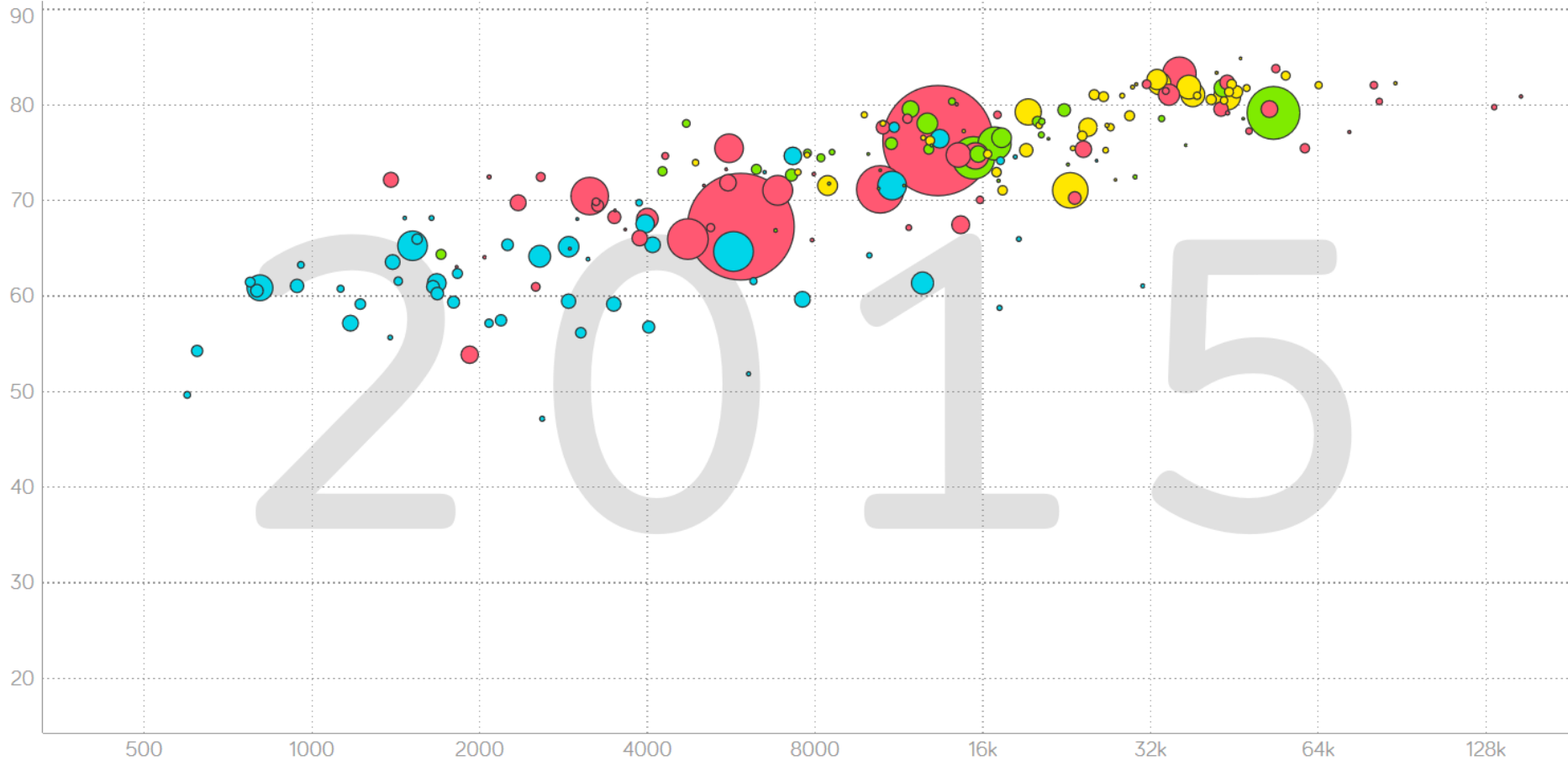
Bubble Charts

- Trivariate Data
 - Three quantitative variables
- Do not allow for accurate comparison.
- Colors can be used to show different categories.



Bubble Charts

Life expectancy, years ?



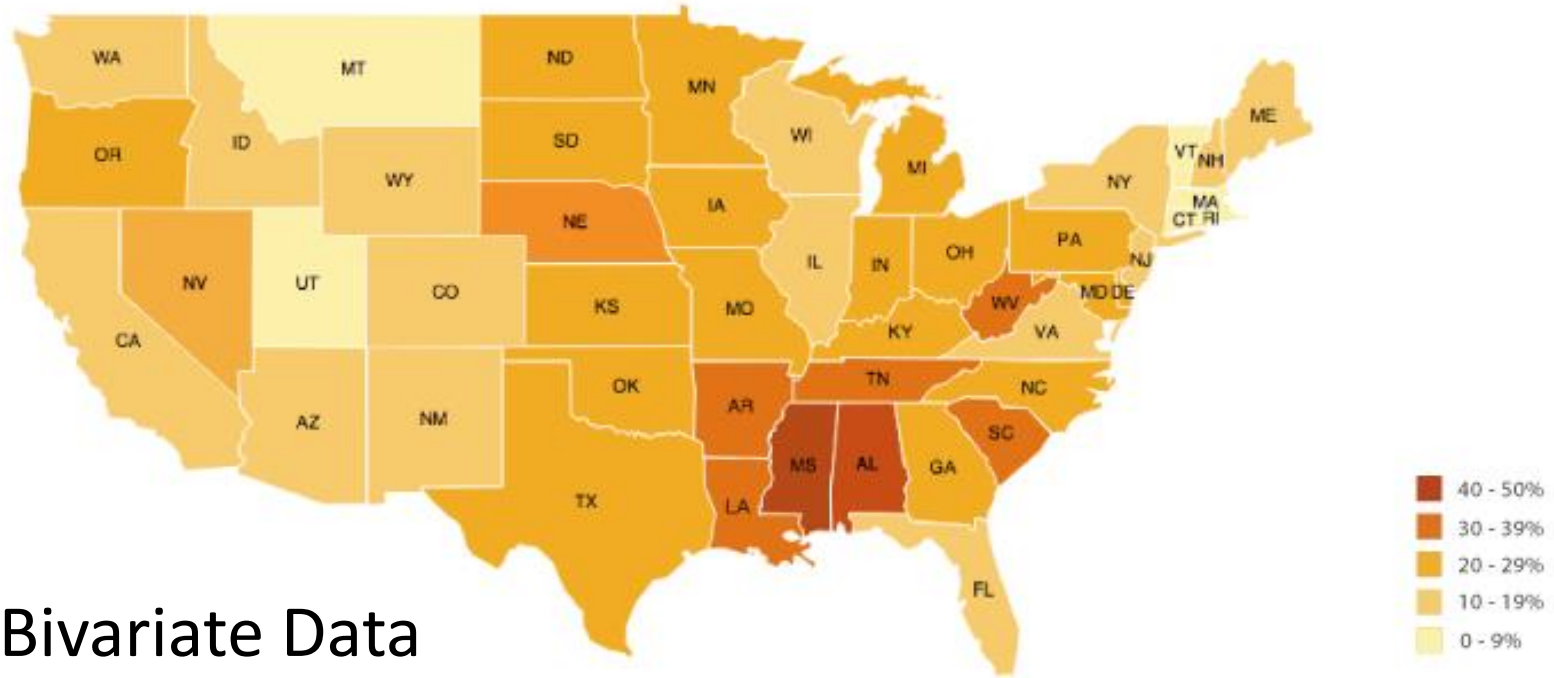
Income per person, GDP/capita in \$/year adjusted for inflation & prices ?

▲ DATA DOUBTS



Figure from <http://gapminder.com>

Chloropleth Map



- Bivariate Data
 - Quantitative value over a geographical areas/regions
- Data are coloured, shaded or patterned in different ways.
- Good for an overview, not for accurate comparison.
- Small areas can be underemphasized.

Word Cloud

- Word Clouds displays how frequently words appear in a given body of text, by making the *size of each word proportional to its frequency*.
- Arrangement and color can vary a lot.
- Word Clouds can be used to compare two bodies of text or to give a quick idea of repeating keywords (e.g. used by researchers to summarize the content of their papers).

Word Clouds

- Disadvantages:
 - Long words are emphasized over short words.
 - Words whose letters contain many ascenders and descenders may receive more attention.
 - No accuracy comparison, mainly used for aesthetic reasons.

Word Clouds (example)



Word Clouds (example)



Word Clouds (example)



Summary

- Visualization plays a fundamental role in data analysis.
- Information coming from data. Data should be collected and processed properly.
- Depending on goal and data type, the use of certain graphs is more effective than others.

Questions ?