

PCC3110

Tratamento de dados

Prof. Dr. Cheng Liang Yee

Prof. Dr. Fernando Akira Kurokawa

Prof. Dr. Sérgio Leal Ferreira

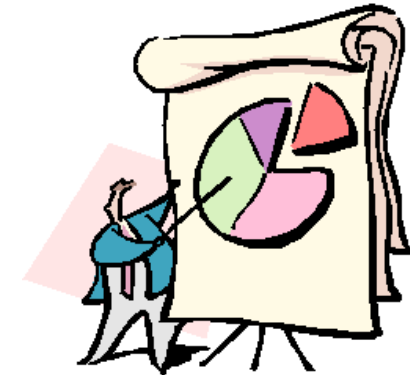
Fases do Processo Metodológico



Formulação do problema



Coleta dos dados



Conclusões e generalizações



Formulação da hipótese



Análise dos dados



Redação

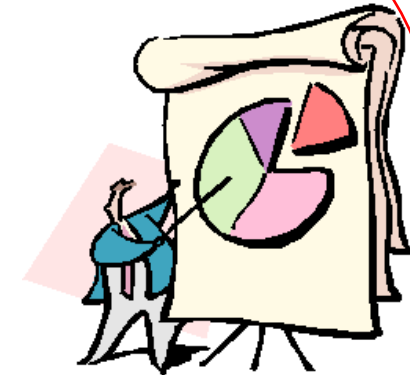
Fases do Processo Metodológico



Formulação do problema



Coleta dos dados



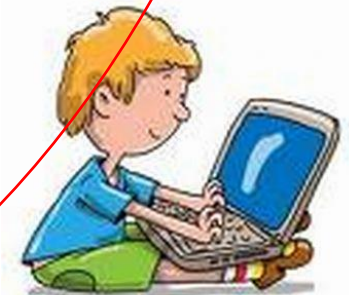
Conclusões e generalizações



Formulação da hipótese



Análise dos dados

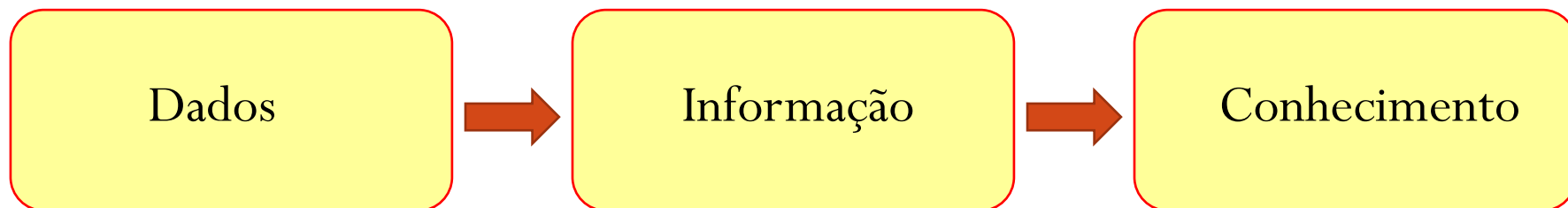


Redação

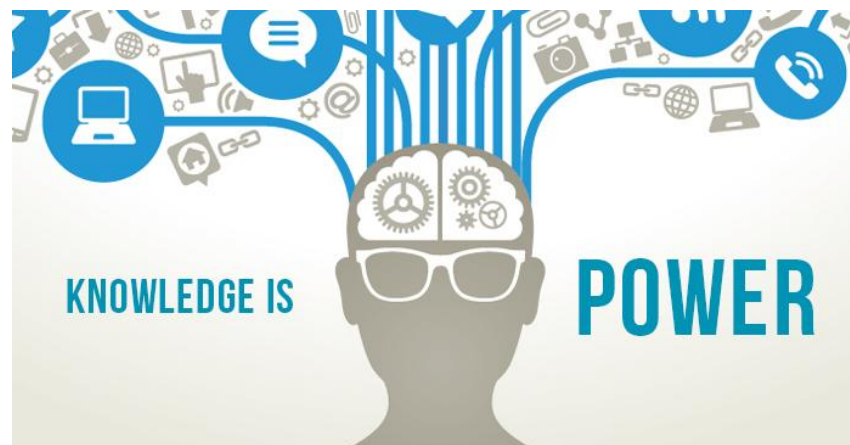
Tratamento, análise dos dados e apresentação dos resultados

Considerações iniciais

Objetivo : Geração de conhecimento



- O processo e a evolução:
Quando mais próximo do conhecimento consegue atingir, maior a contribuição!



Cuidado!

**Os números podem ser manipulados
para revelar ou enganar!**

- Por isso, a importância da:
 - Organização e tratamento
 - Análise
 - Forma de apresentação dos resultados

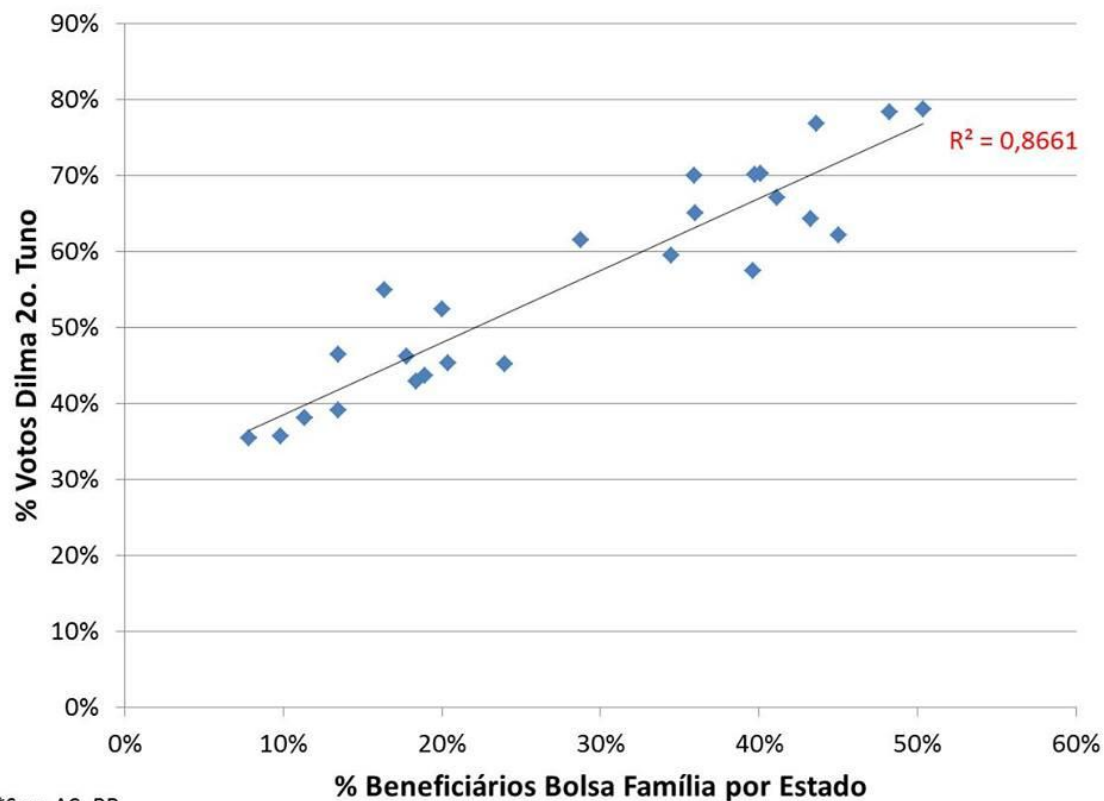
E da forma correta!



Exemplos interessantes

- Uma visão sobre os resultados da eleição presidencial 2014
 - QUANTO VALE A LIBERDADE NO BRASIL?
 - Christian Fleury

A Democracia da Barriga - 2.Turno

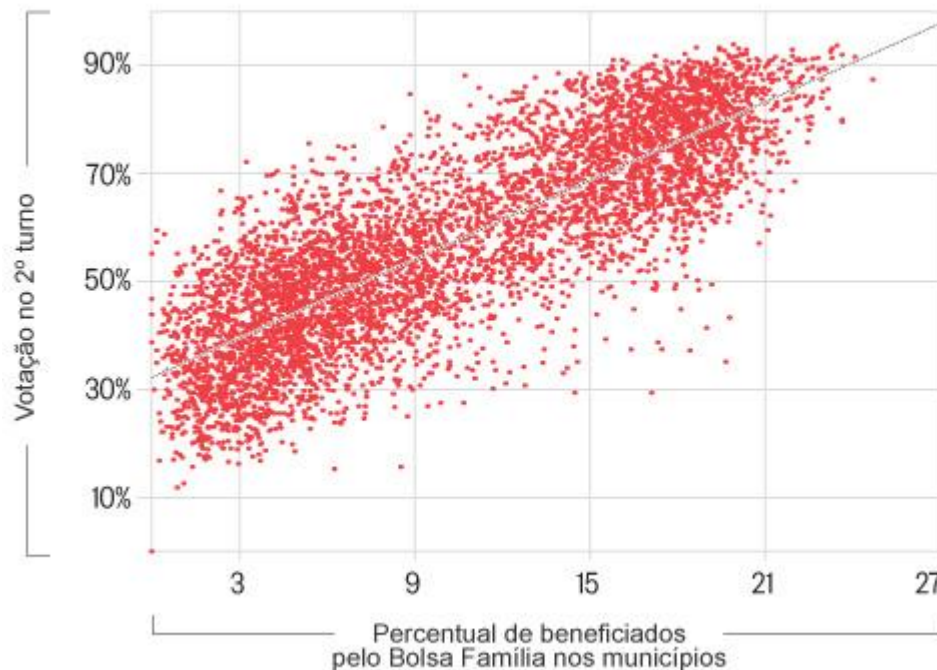


*Sem AC, RR

Exemplos interessantes

- Outra visão sobre os resultados da eleição presidencial 2014
 - Ganho de votos de Dilma no 2º turno não tem relação com Bolsa Família
 - O GLOBO

NO 2º TURNO



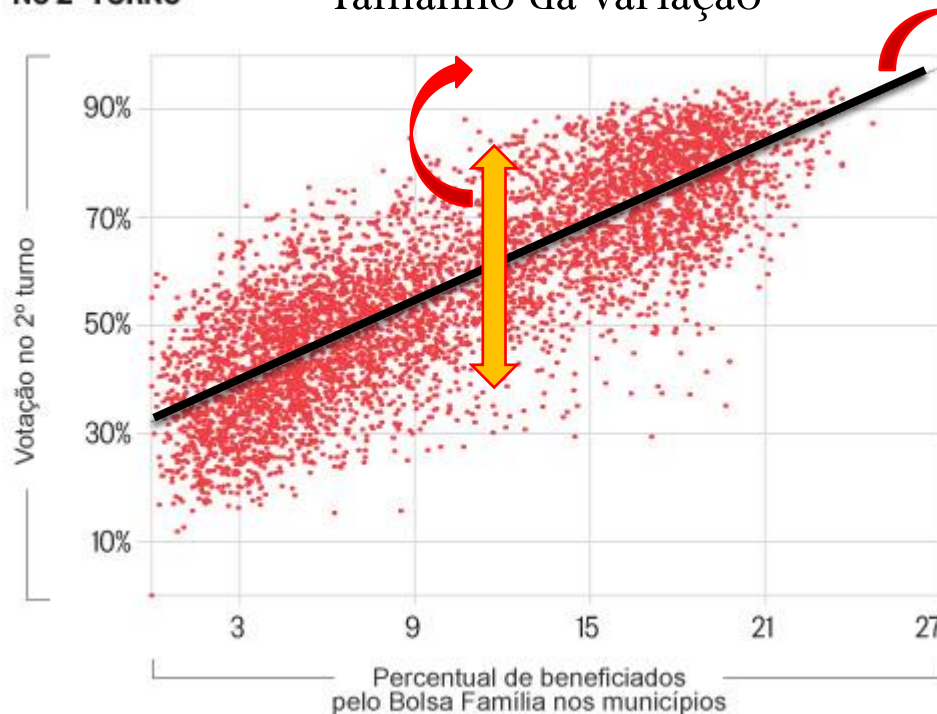
O que você acha?

Exemplos interessantes

- Outra visão sobre os resultados da eleição presidencial 2014
 - Ganho de votos de Dilma no 2º turno não tem relação com Bolsa Família
 - O GLOBO

NO 2º TURNO

Tamanho da variação

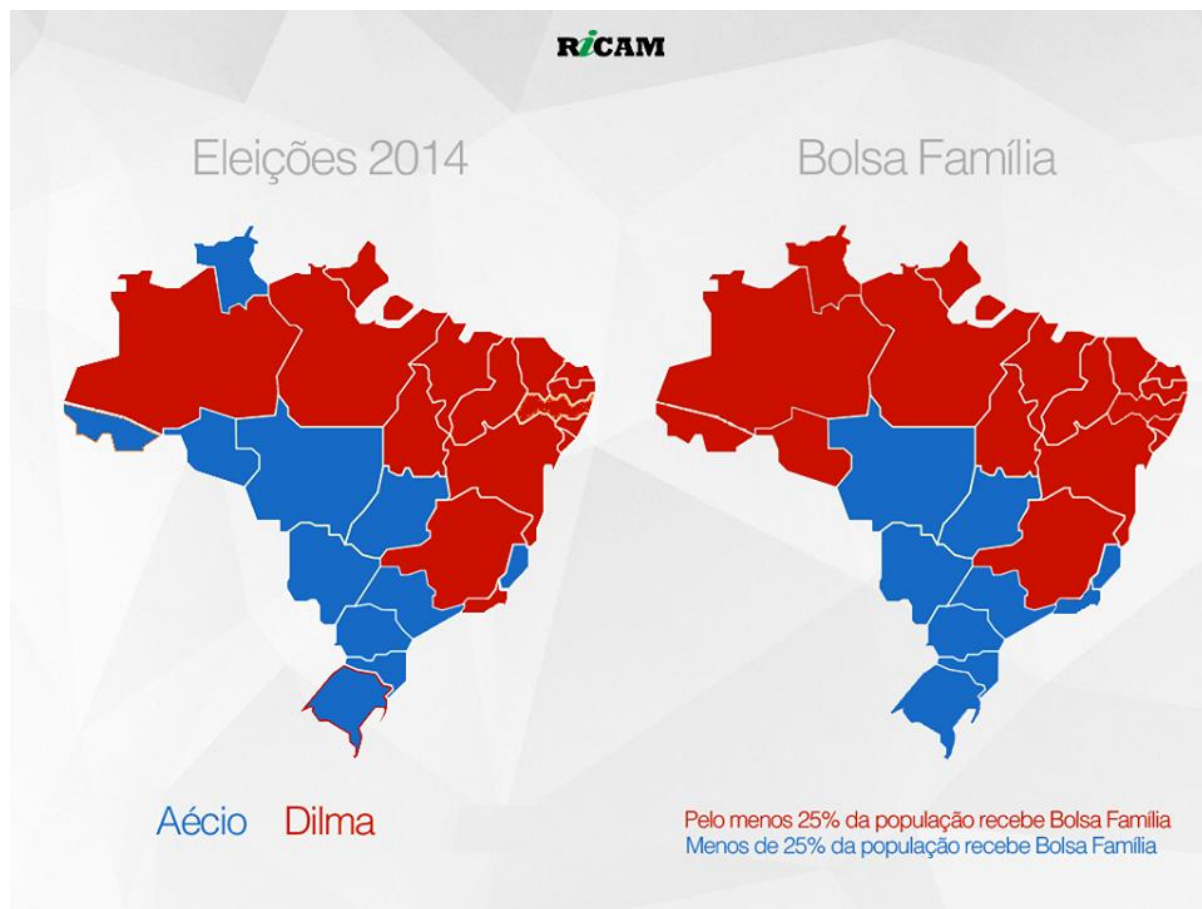


Comportamento geral do grupo

Sob o ponto de vista **estatístico**, o indivíduo não tem interesse e só passa a ser interessante quando faz parte de um todo!

Exemplos interessantes

- Com exceção de AC, RO, RR e RJ, estes dois mapas explicam todo o resultado da eleição presidencial.
- Ricardo Amorim



Exemplos interessantes

- Menos ódio, por favor...

Mapa revela mistura de votos e mostra pouca diferença entre Nordeste e Sudeste

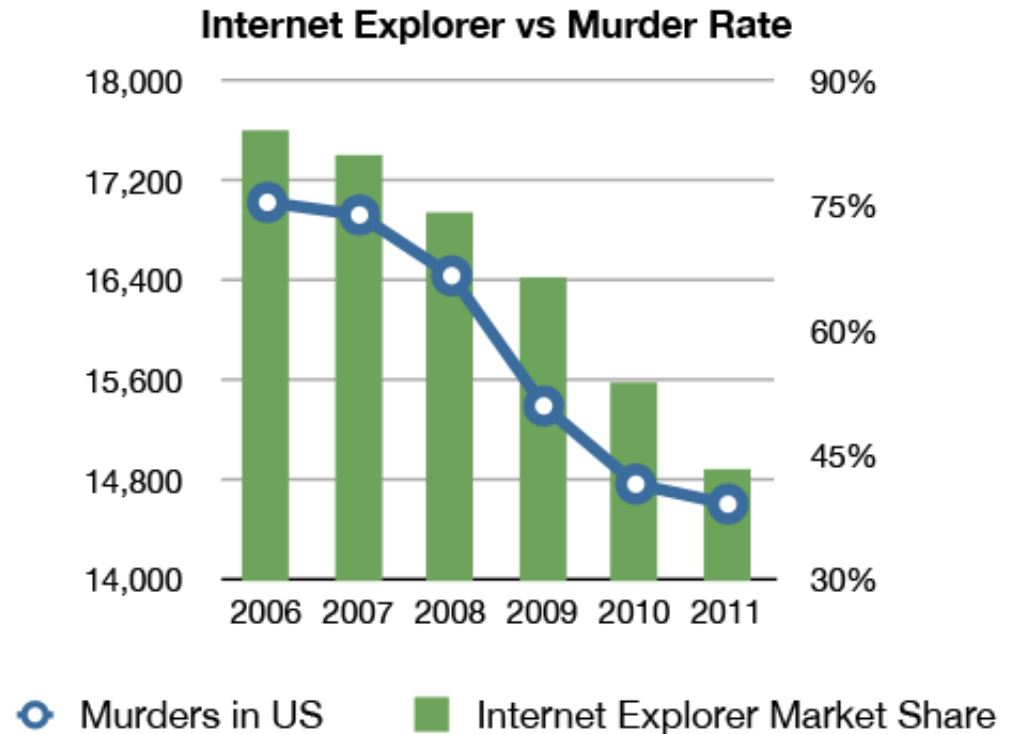
<http://glo.bo/1w8Wvbf>

- Jornal Correio



Exemplos interessantes

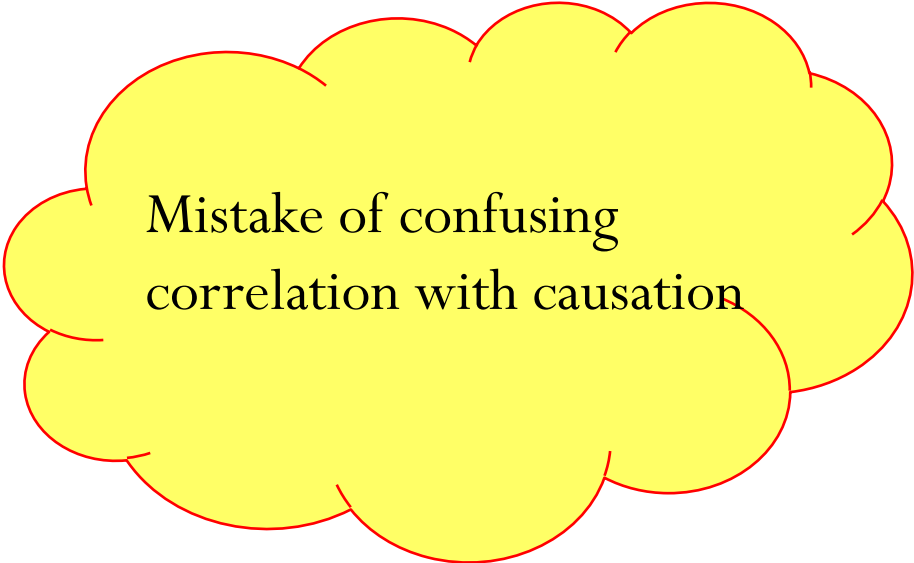
- Estudos apontam... que boa parte dos estudos estão equivocados
- Christian Fleury
 - Esta vai para quem acredita cegamente em qualquer coisa.
 - O gráfico é uma sátira que associa a queda dos assassinatos nos EUA à queda do uso do Internet Explorer, o navegador da microsoft.



- O mundo está tão cheio de informações, que qualquer um pode provar qualquer coisa juntando dados de origens diferentes.
- Não é difícil encontrar falsas correlações entre dados que caminham em sentidos semelhantes, porém motivados por causas completamente diferentes.

Leitura recomendada

- Clearing up confusion between correlation and causation.
- <https://theconversation.com/clearing-up-confusion-between-correlation-and-causation-30761>



Mistake of confusing correlation with causation

- What do we actually mean by research and how does it help inform our understanding of things? Today we look at the **dangers of making a link between unrelated results.**

Como estabelecer relação causa-efeito?

- Existência da correlação
- Sequencia dos acontecimentos (ordem temporal)
- Inexistência de outros fatores causadores

“One must always be wary when drawing conclusions from data! “

- “In order to establish cause-and-effect, we need to go beyond the statistics and look for separate evidence (of a scientific or historical nature) and logical reasoning. “
- “Correlation may prompt us to go looking for such evidence in the first place, but it is by no means a proof in its own right.”

Tipos de relação causa-efeito

- Determinista

Se ocorrer x , sempre ocorrerá y

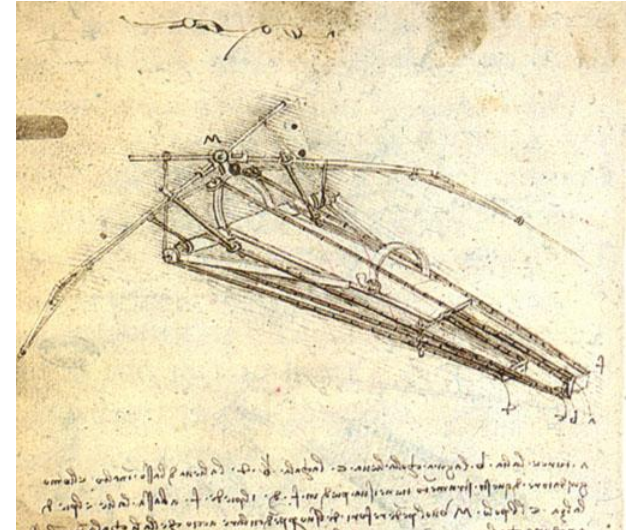
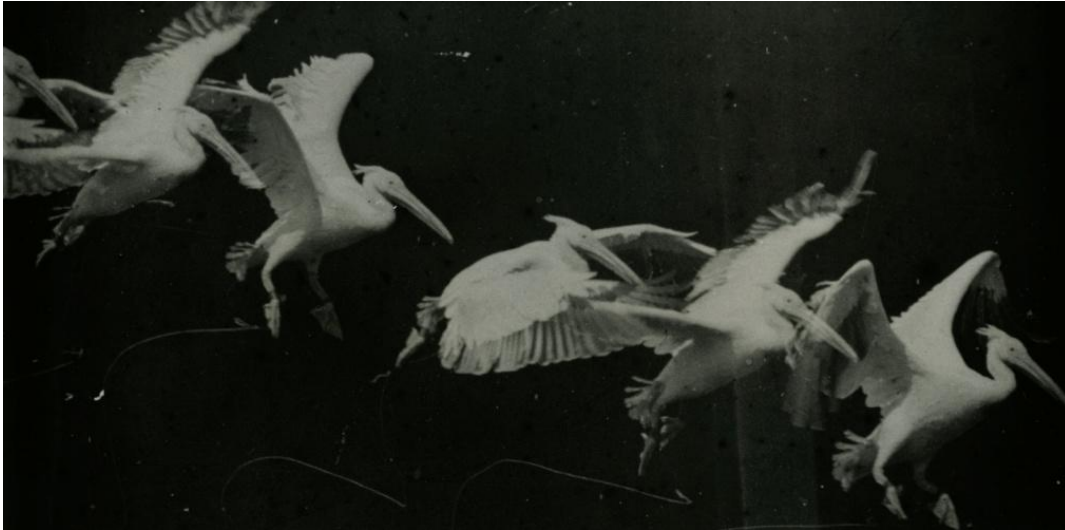
- Suficiente

x causa y

- Probabilística ou estocástica

Se ocorre x , há possibilidade de ocorrer y

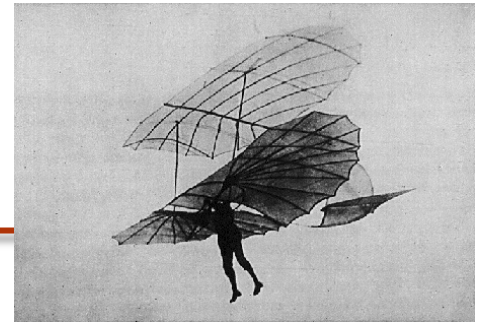
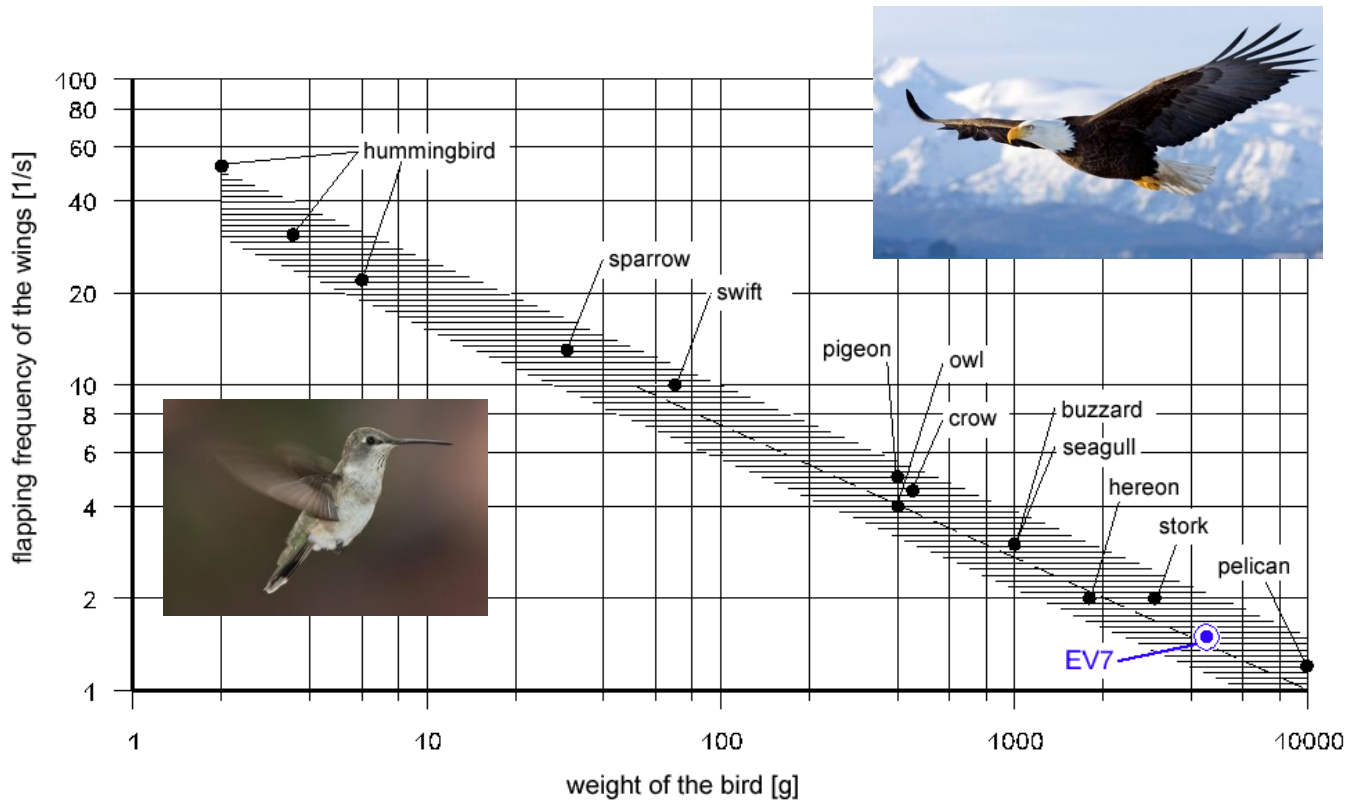
Exemplos interessantes



- Os pioneiros da aviação, que tentaram voar imitando os pássaros, batendo as asas...

<http://www.youtube.com/watch?v=gN-ZktmjIfE>

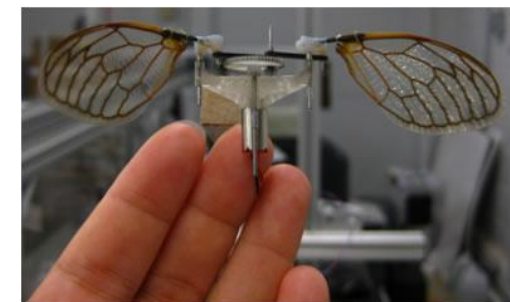
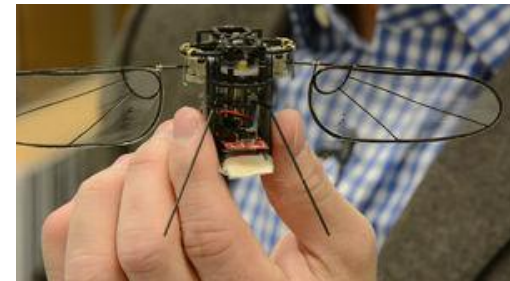
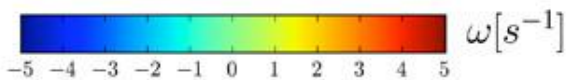
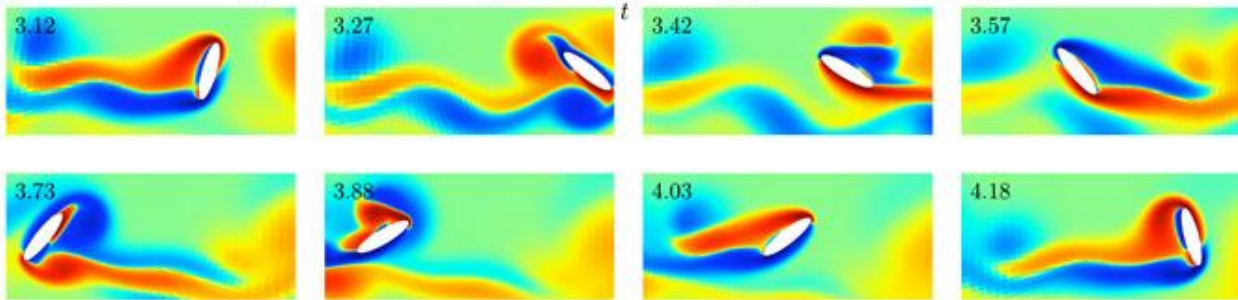
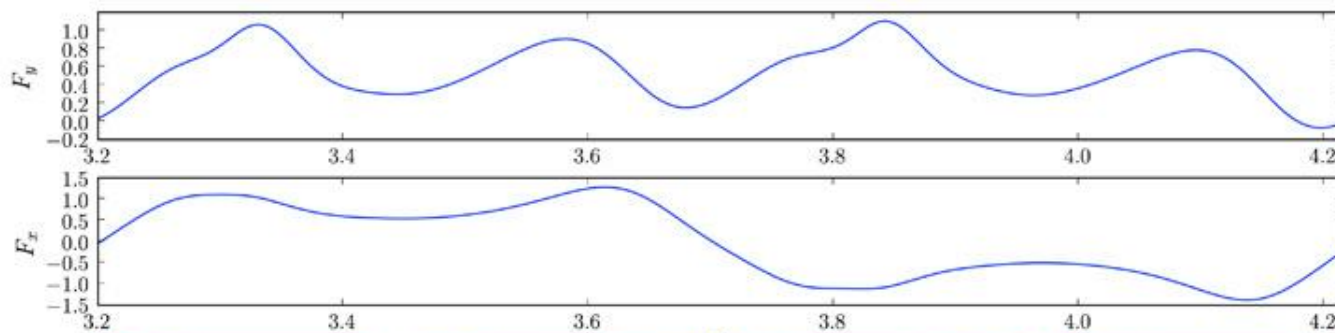
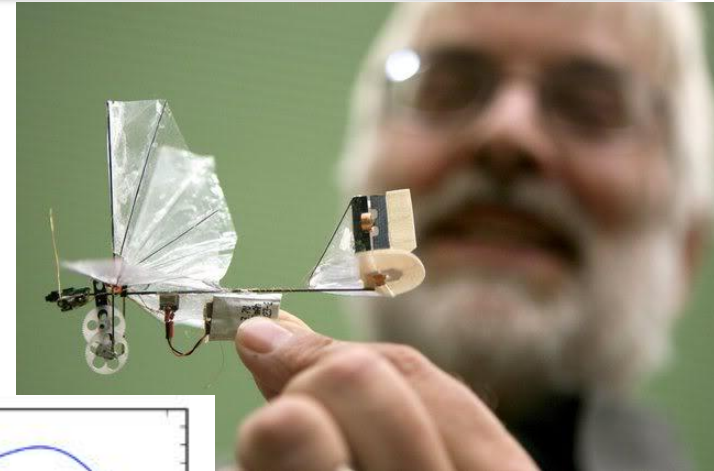
Exemplos interessantes



- Se soubessem disso antes... Teriam estendidos as asas e sair voando sem batê-las...

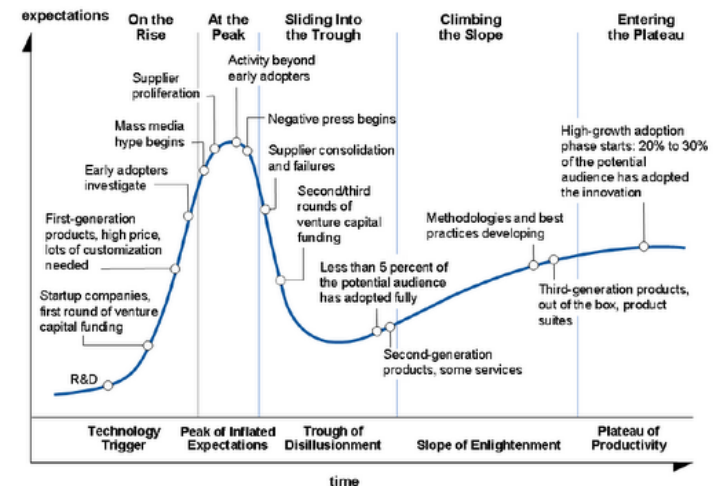
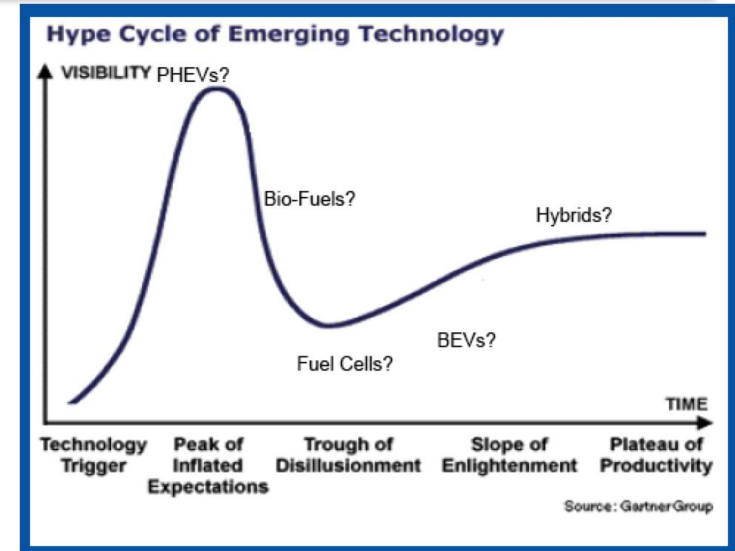
A onda do ornitóptero

- Mas sempre tem futuro!!!
E jamais devemos desistir!!!



Gartner Hype Cycle

- Hype Cycle foi proposto pela empresa de pesquisa e consultoria Gartner para representar graficamente a maturidade, a adoção e a aplicação social de uma tecnologia específica.

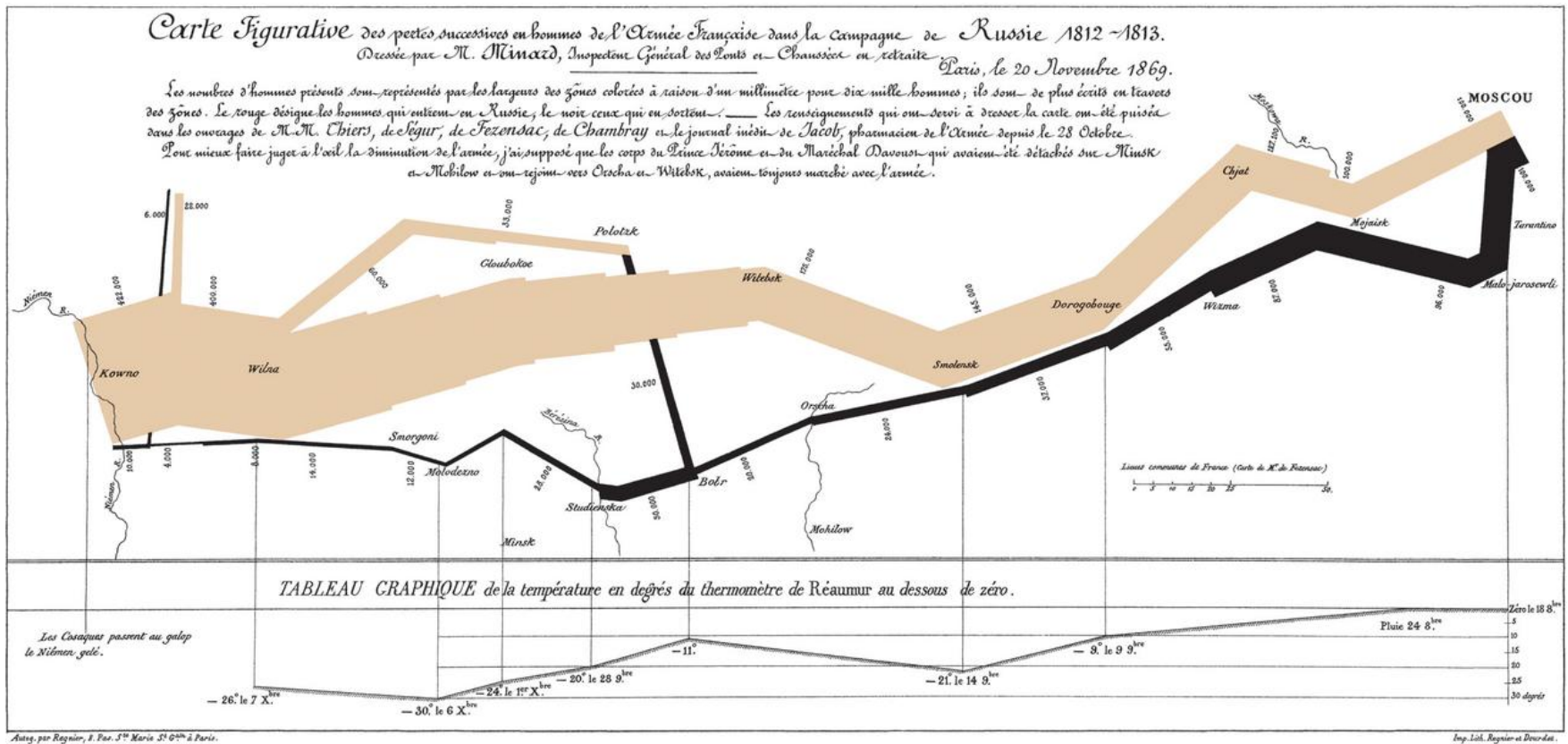


Source: Gartner (August 2010)

Modern information scientists say the illustration may be the best statistical graphic ever drawn.

Exemplos interessantes

- Mapa do Charles Minard sobre a desastrosa campanha da Rússia do Napoleão, em 1812.
- O gráfico é notório por representar em 6 dados em 2D: Número da tropa Napoleônica, distância, temperatura, latitude, longitude, direção do deslocamento e posição com datas.



Investigação estatística

- Envolve, de um modo geral, quatro fases:
 - **Formulação do problema** a investigar, na forma de questões que se procuram responder através de dados;
 - Planejamento adequado para **coleta de dados** apropriados;
 - **Organização e tratamento dos dados** coletados, através de tabelas, gráficos e algumas medidas;
 - **Interpretação** dos resultados obtidos **e formulação de conclusões**.

Organização e tratamento dos dados

- “Limpar” os dados:
 - É comum, quando se procede a uma análise de dados coletados verificar que estes contêm erros, acidentais ou não acidentais.
 - Assim, antes de se proceder ao tratamento dos dados através de tabelas, gráficos ou do cálculo de medidas, deve-se olhar criticamente para os dados coletados, com o objetivo de os “limpar” dos erros.
- Exemplo: Na colata da medida dos pés, se obtiver a informação de 300 cm, obviamente que este valor está errado...
- Formas de representação dos dados dependem da natureza das variáveis em jogo (qualitativas, nominais ou ordinais, e quantitativas, discretas ou contínuas), e também para alguns aspectos que facilmente induzem em erro.

Natureza das variáveis

- Recapturando....

Quantitativas

Contínuas

Discretas

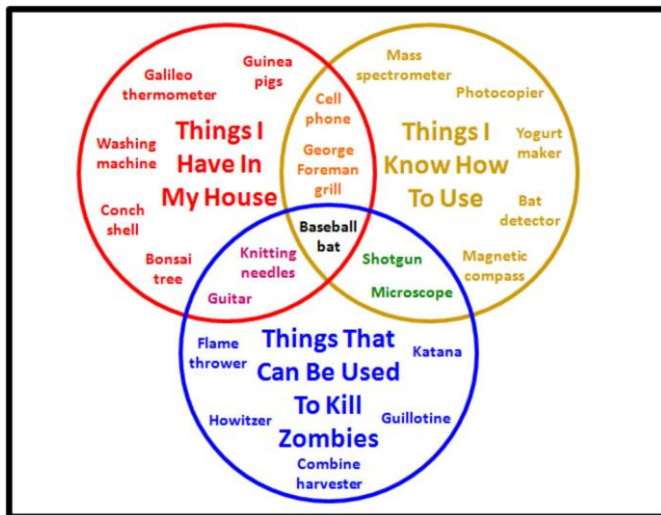
Qualitativas

Categoria ordinal

Categoria nominal

Diagramas

- Diagramas de Venn: utilizam círculos ou retângulos para uma classificação rápida de objetos ou números, que partilhem características comuns.
- Diagramas de Carroll são tabelas retangulares para organizar dados ou objetos segundo critérios de sim/não.



	Prime	Not prime
Even	2	4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26
Not even	3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41	1, 9, 15, 21, 25, 27, 33, 35, 39, 41, 45, 49



Tabelas e gráficos para dados qualitativos

- Esquemas de contagem gráfica (*tally charts*): uma representação muito simples que se podem construir diretamente a partir do conjunto de dados ou durante o processo de recolha.
- Tabela de frequências para dados qualitativos: reflete a forma da distribuição da variável em estudo, na amostra considerada, isto é, quais as categorias ou modalidades que assume, assim como a frequência (absoluta e/ou relativa) com que assume essas modalidades.

Favorite Color

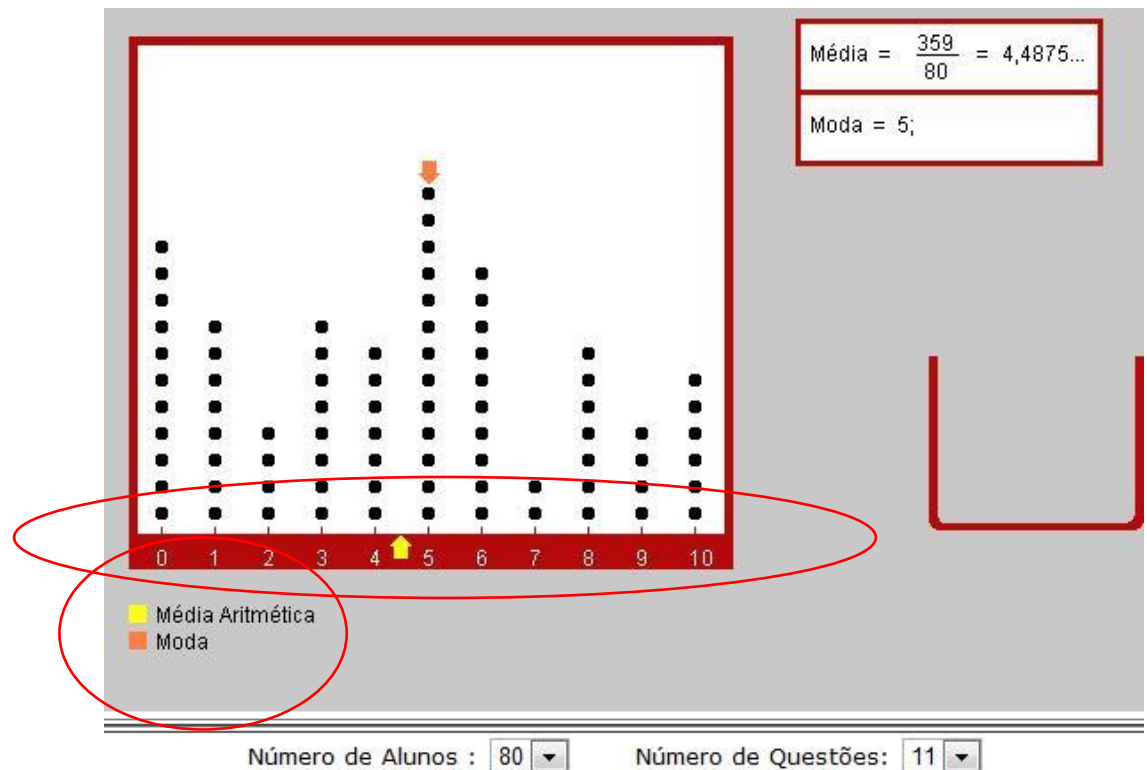
Color	Tally	Frequency
red		5
orange		3
yellow		2
green		3
blue		5
indigo		2
violet		1



"I've completely lost track. Is it 1 million or 2 million notches to A.D.?"

Tabelas e gráficos para dados qualitativos

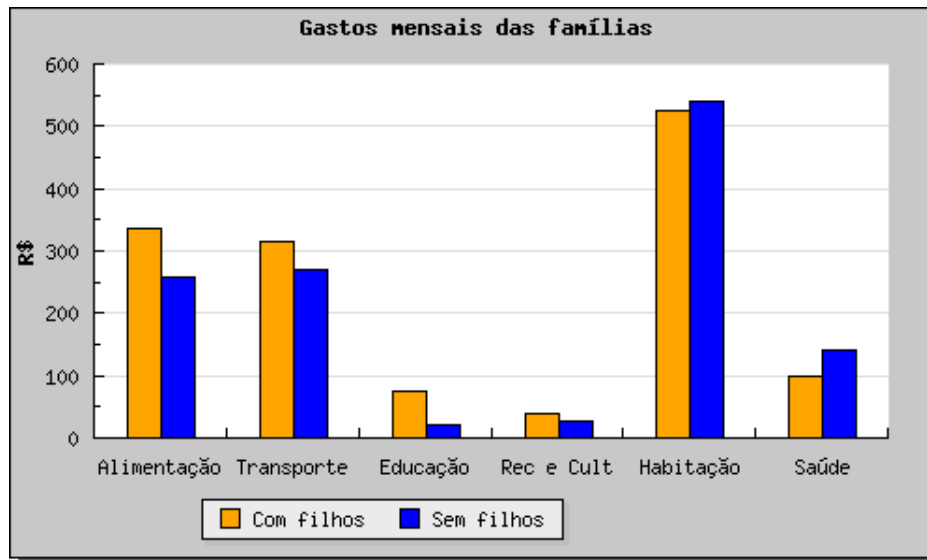
- Gráfico de pontos: representação gráfica mais simples que se pode obter, que não necessita de nenhuma organização prévia dos dados, e pode ir construindo, à medida que se recolhem os dados.



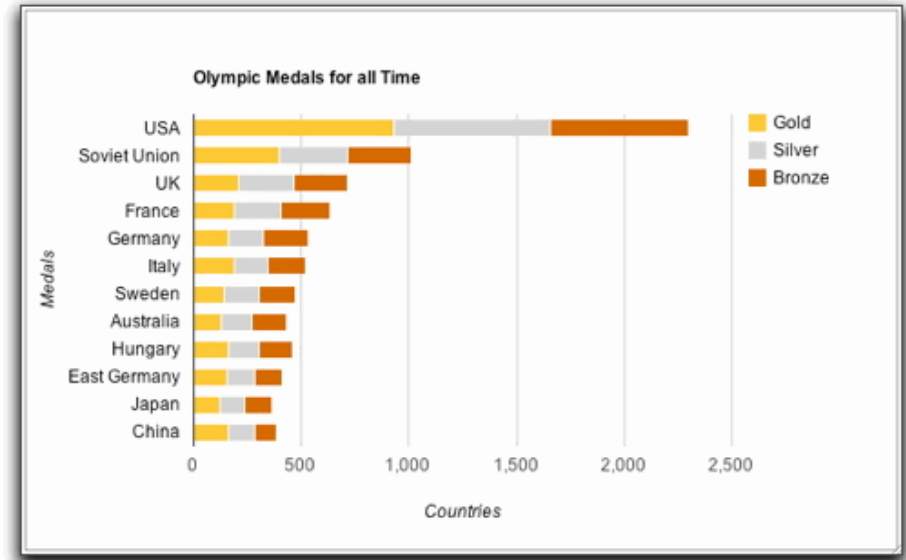
Tabelas e gráficos para dados qualitativos

- Gráfico de barras: representação gráfica da informação de uma tabela de frequências.
- A ordem das categorias é:
 - arbitrária para dados qualitativos nominais e,
 - sequencial para a dados qualitativos ordinais.

Comparativo de 2 conjunto de dados lado ao lado

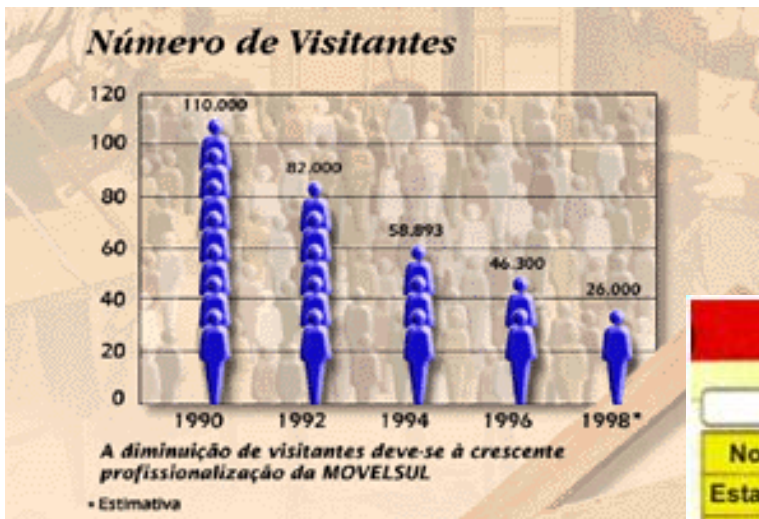


Comparativo de 3 conjunto na mesma barra



Tabelas e gráficos para dados qualitativos

- Pictograma: uma representação gráfica que usa símbolos alegóricos às variáveis que se estão a estudar.



O número do símbolo
proporcional à
quantidade

O tamanho do símbolo
proporcional à
quantidade



Tabelas e gráficos para dados qualitativos

- Gráfico circular: a base desta representação é um círculo que representa a forma como o total de um conjunto de dados se distribui pelas categorias.

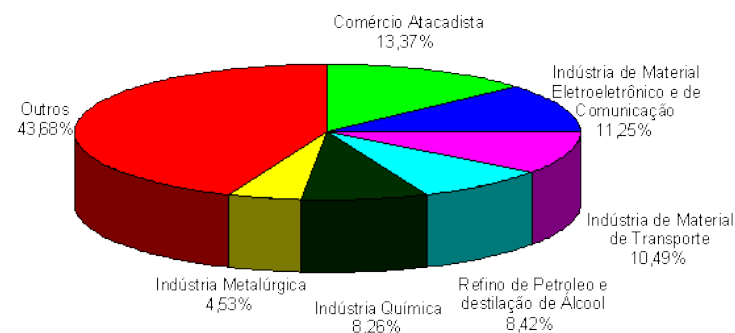
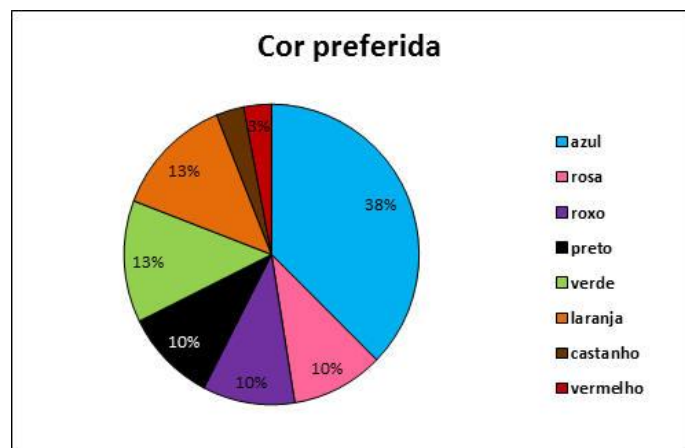
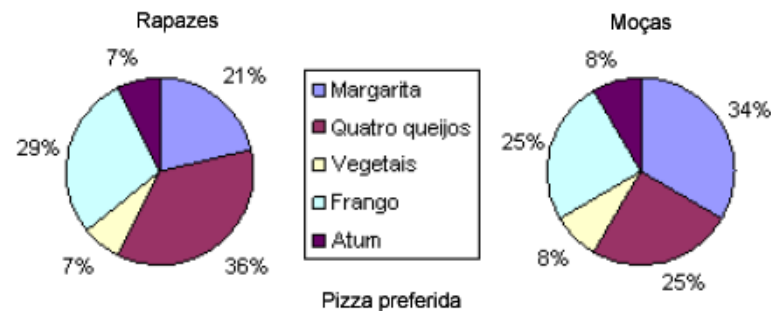


Gráfico circular é uma representação por excelência, no entanto, gráfico de barras é mais claro quando a distribuição apresenta muitas categorias ou quando os valores das frequências de algumas das categorias estão próximos.

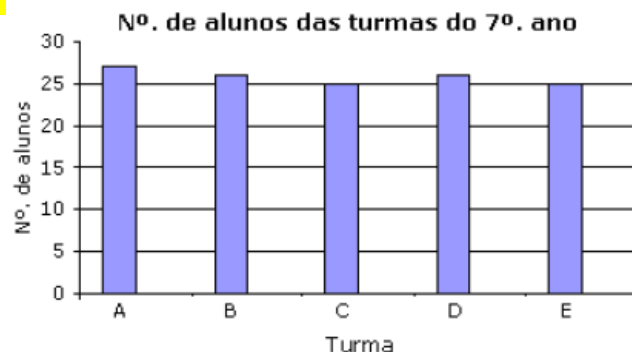


Frequência x dados

- É comum utilizarem-se gráficos de barras para representar os próprios dados e não as frequências com que as diferentes classes ou categorias surgem no conjunto de dados.

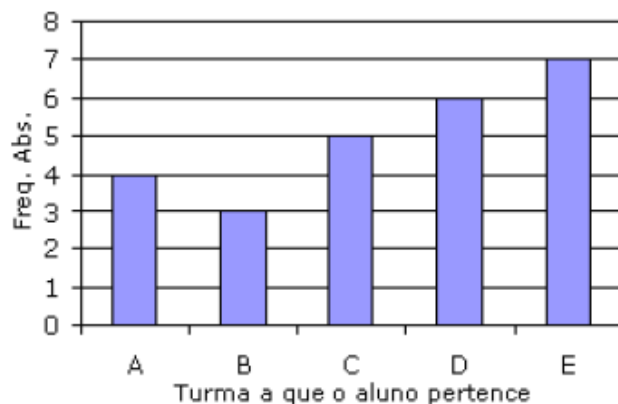
Não é uma tabela de frequências!

Turma	Nº. de alunos
A	27
B	26
C	25
D	26
E	25



É uma tabela de frequências!

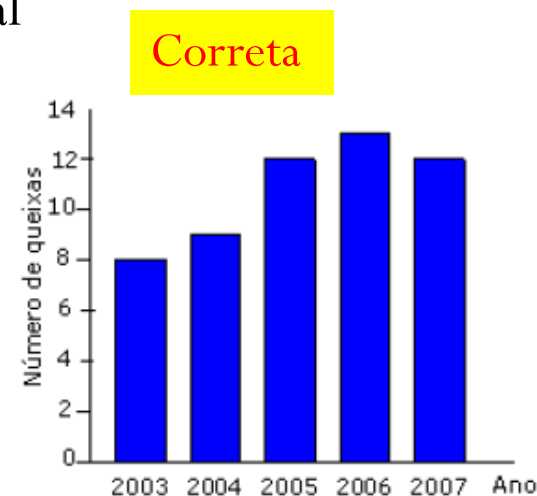
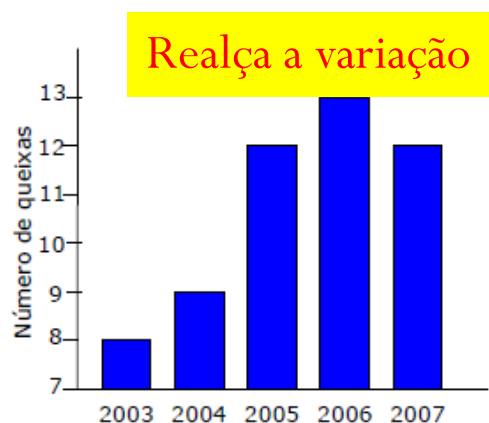
Turma	Freq. Absoluta
A	4
B	3
C	5
D	6
E	7
Total	25



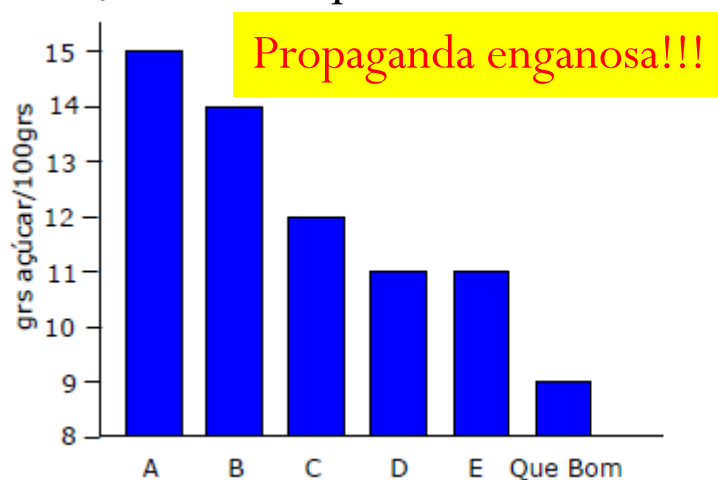
- Em Estatística um gráfico de barras deve representar somente as frequências absolutas ou relativas de um conjunto de dados.

Atenção às escalas!

- Exemplo: número de queixas recebidas num hospital



- Exemplo: teor de açúcar num produto



Tabelas e gráficos para dados quantitativos discretos

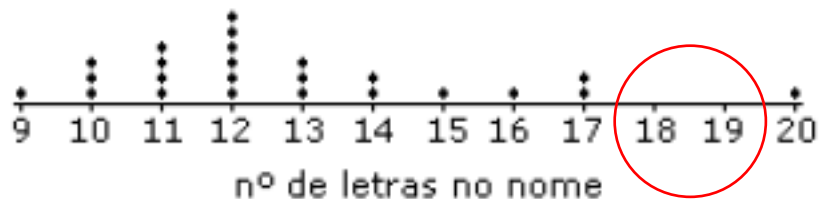
- Tabela de frequências para dados quantitativos discretos:
 - Sua construção é idêntica à construída para dados qualitativos,
 - Considera-se agora para classes os valores distintos que surgem no conjunto de dados.
 - Adicionar as colunas de frequências acumuladas.

Número de irmãos x_i	Frequência absoluta n_i	Frequência relativa f_i	Frequência abs. acumulada N_i	Frequência rel. acumulada F_i
0	2	10%	2	10%
1	6	30%	8	40%
2	6	30%	14	70%
3	4	20%	18	90%
4	2	10%	20	100%
Total	20	100%		

- No caso das variáveis qualitativas, só tem sentido calcular as frequências acumuladas se forem ordinais.

Gráfico de pontos e gráfico de barras para dados quantitativos discretos

- Gráfico de pontos e gráficos de barras: semelhante àquele para os dados qualitativos.



Deve marcar-se no eixo a sequência completa dos valores entre o mínimo e o máximo observados, mesmo que alguns desses valores não constem da amostra.

N.º de letras no nome x_i^*	Freq. Abs. n_i	Freq. Rel. f_i
9	1	0,042
10	3	0,125
11	4	0,167
12	6	0,250
13	3	0,125
14	2	0,083
15	2	0,042
16	1	0,042
17	2	0,083
20	1	0,042
Total	24	1,000

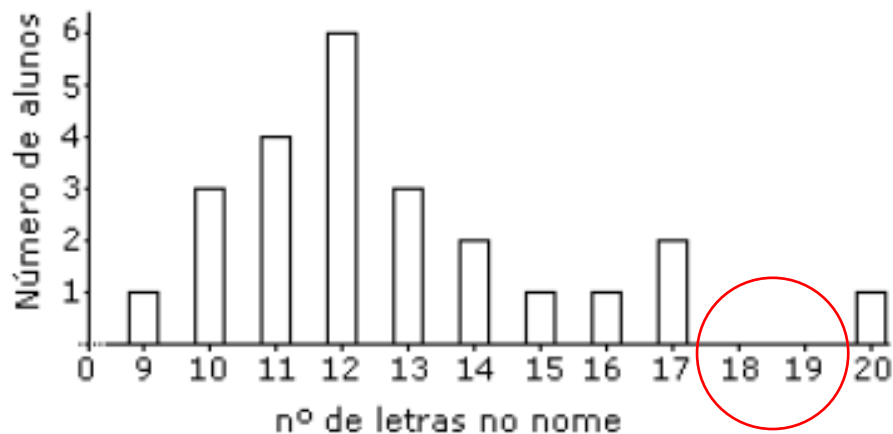
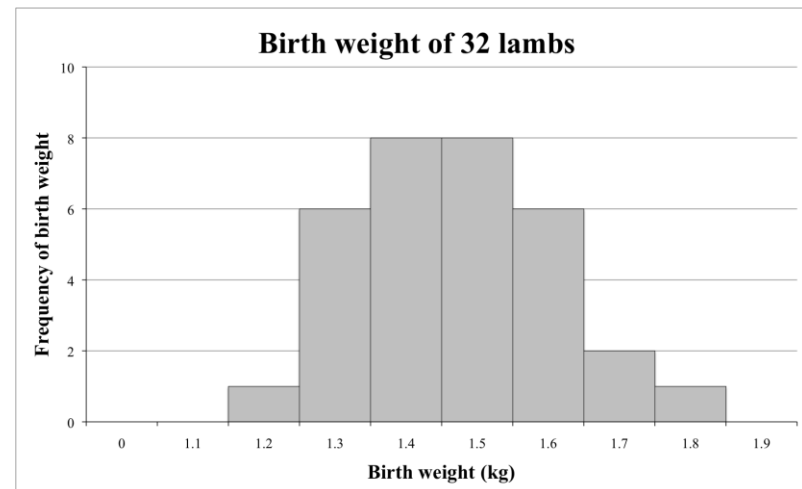


Gráfico de pontos e gráfico de barras para dados quantitativos discretos

- Quanto tem um número demasiado de classes, para evitar tabela de frequência grande, pode **agrupar as classes em intervalos**.
- A representação gráfica para os dados organizados desta forma já não pode ser um diagrama de barras, pois não existe um ponto onde colocar a barra, uma vez que as classes são intervalos.
- Neste caso, a representação gráfica adequada é o **histograma**.



Tabelas e gráficos para dados quantitativos contínuos

- Para dados quantitativos contínuos, o número de valores distintos é demasiadamente grande:
 - A metodologia utilizada para construir as tabelas de frequências de dados quantitativos discretos não pode ser aqui utilizada, pois corre o risco de a frequência observada para cada valor distinto ser 1!
 - Alternativa é **considerar classes na forma de intervalos**.

Deve marcar no eixo a sequência completa dos valores entre o mínimo e o máximo observados, mesmo que alguns desses valores não constem da amostra.

Tabela 1 - Idades dos alunos de estatística (2011)

Idades	xi	fi	fr	fr%	Fi	Fr%
18 ----23	20,5	44	0,80	80,00%	44	80,00%
23 ----28	25,5	7	0,13	12,73%	51	92,73%
28 ----33	30,5	0	0,00	0,00%	51	92,73%
33 ----38	35,5	2	0,04	3,64%	53	96,36%
38 ----43	40,5	1	0,02	1,82%	54	98,18%
43 ----48	45,5	0	0,00	0,00%	54	98,18%
48 ----53	50,5	1	0,02	1,82%	55	100,00%
Total	...	55	1	100,00%

Regra de Sturges

- O primeiro passo no processo de agrupamento dos dados é saber em **quantas classes** vamos agrupar os dados.
- Regra de Sturges:

Organizar uma amostra, de dados contínuos, de dimensão n , pode considerar-se para número de classes o valor k , onde k é o menor inteiro tal que $2^k > n$.

- Exemplo: se o número de elementos da amostra for 50, como nos exemplos apresentados anteriormente, o número aconselhado de classes é 6, já que $2^5 < 50$ e $2^6 > 50$.

Tabela de frequências para dados quantitativos contínuos

- Para dados quantitativos contínuos, adiciona uma coluna de “Representantes da classe”.
- Exemplo: altura de um aluno da escola

Classes	Representante da Classe x'_i	Freq. Abs. n_i	Freq. Rel. f_i	Freq. Abs. Acum	Freq. Rel. Acum.	Freq. Rel. Acum. (%)
[130, 135[132,5	7	0,14	7	0,14	14
[135, 140[137,5	9	0,18	16	0,32	32
[140, 145[142,5	11	0,22	27	0,54	54
[145, 150[147,5	14	0,28	41	0,82	82
[150, 155[152,5	5	0,10	46	0,92	92
[155, 160[157,5	4	0,08	50	1,00	100
Total		50	1,00			

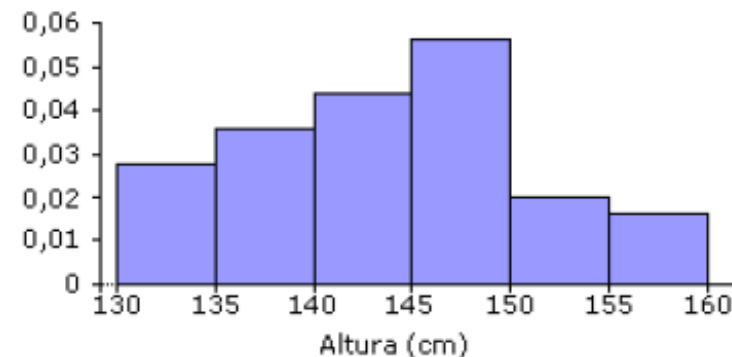
Usualmente, o ponto médio, x'_i , do intervalo de classe.

Histograma

- Gráfico formado por uma sucessão de retângulos adjacentes, tendo cada um por base um intervalo de classe e com área igual (ou proporcional) à frequência relativa (ou absoluta) dessa classe.
- **Ao contrário do gráfico de barras**, em que estas estão separadas e em que o que é relevante é a altura de cada uma, **no histograma as barras (retângulos) estão juntas e o que é importante é a área de cada uma.**
- Considerando então para áreas das barras as frequências relativas, vemos que **a área total ocupada pelo histograma é igual a 1 ou 100%.**

É conveniente acrescentar uma nova coluna com as frequências relativas dividida pela amplitude de classe. Os valores desta coluna serão as alturas dos retângulos.

Classes	Rep. Classe x'_i	Freq. Abs. n_i	Freq. Rel. f_i	Altura rectângulo classe $i=f_i/h$
[130, 135[132,5	7	0,14	0,028
[135, 140[137,5	9	0,18	0,036
[140, 145[142,5	11	0,22	0,044
[145, 150[147,5	14	0,28	0,056
[150, 155[152,5	5	0,10	0,020
[155, 160[157,5	4	0,08	0,016
Total		50	1,00	



Histograma

- Exemplo com intervalos não uniformes:
 - a área de cada barra continua igual (ou proporcional) à frequência relativa (ou absoluta) dessa classe.
 - a área total ocupada pelo histograma continua sendo igual a 1 ou 100%.

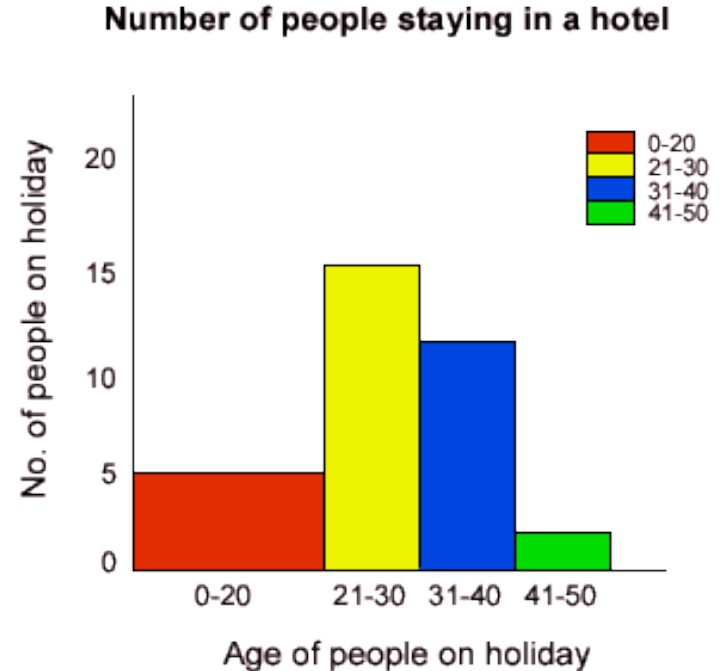


Diagrama de caule-e-folhas

- Representação que se situa entre a tabela e o gráfico. Apresenta os verdadeiros valores da amostra, mas de uma forma que faz lembrar o histograma.
- Exemplos: tempo para execução de uma tarefa:
- 59, 38, 47, 23, 48, 55, 37, 48, 53, 37, 52, 39, 54, 57, 38, 46, 40, 41, 62, 63, 38, 65, 44, 68, 27, 35, 46, 60.

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

- Cuidado na escolha dos dígito(s) dominantes.
- Dá uma informação visual sobre a forma como os dados estão distribuídos.
- Muito útil para ordenar rapidamente a amostra.
- Muito sugestiva para comparar duas amostras.
- Facilita o cálculo da mediana e dos quartis.

Parâmetros interessantes

- **Mediana (Me)**: valor do elemento do meio (ou a média dos dois elementos do meio).
- Pelo mesmo processo, as medianas de cada uma das partes, em que os dados ficam divididos pela mediana, obtemos os **quartis**, respectivamente 1.º quartil se for o da parte inferior e 3.º quartil se for da parte superior.

2	3 7	1.º quartil (Q ₁)	
3	5 7 7 8 8 8 9		
4	0 1 4 6 6 7 8 8	Mediana (Me)	
5	2 3 4 5 7 9		
6	0 2 3 5 8	3.º quartil (Q ₃)	

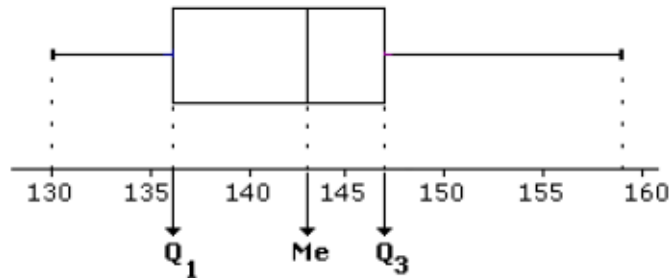
- Utilização do caule-e-folhas para comparar 2 amostras

7	4.							
3	5*							
9 9 8 6	5.							
3 3 0	6*	2	2	2	4			
9 7 6 5	6.	6						
4 3	7*	1	1	1	1	1	4	4
9 7 6	7.	5	5	5	6	6	7	8 9 9
	8*	1	2	3	4			
9 7 7 6	8.	5	6	7	8			
4 4 3 0	9*							
9 6	9.	5						
1 0	10*							

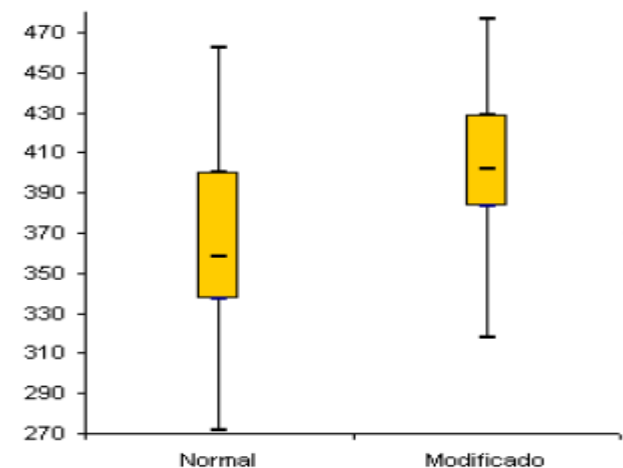
Nesta exemplo, cada dígito dominante é dividido em 2 intervalos iguais representados por sufixos "*" e "."

Diagrama de extremos e quartis

- Representação gráfica muito simples que evidencia de uma forma extremamente eficaz a **mediana, os quartis, o mínimo e o máximo**.

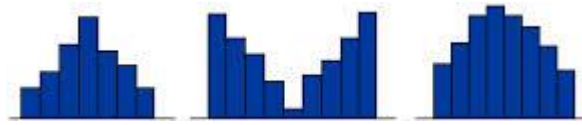


- **Pontos relevantes** ao comparar várias distribuições de dados:
 - Forma da distribuição;
 - Simetria ou ausência de simetria;
 - Variabilidade apresentada.
- Os diagramas de extremos e quartis são úteis para:
 - Comparação das medianas;
 - Comparação da dispersão entre os dados;
 - Identificação de possíveis “outliers”.

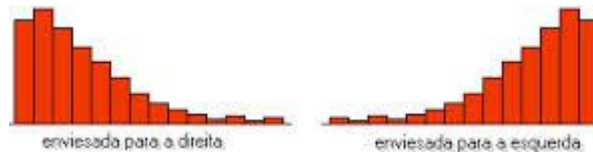


Distribuições de dados

- Distribuições simétricas



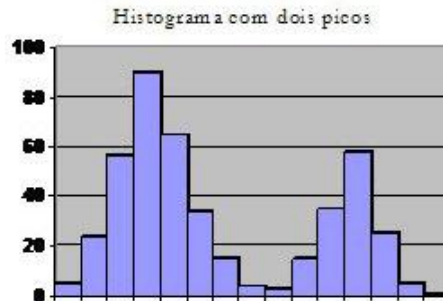
- Distribuições enviesadas



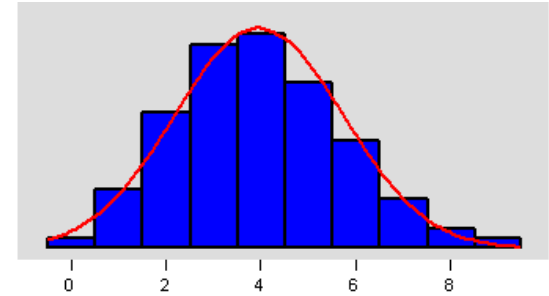
- Distribuições com caudas longas (*Long Tail*)



- Distribuições com vários "picos" ou modas



Aproximação por distribuição normal



Long Tail na análise do mercado

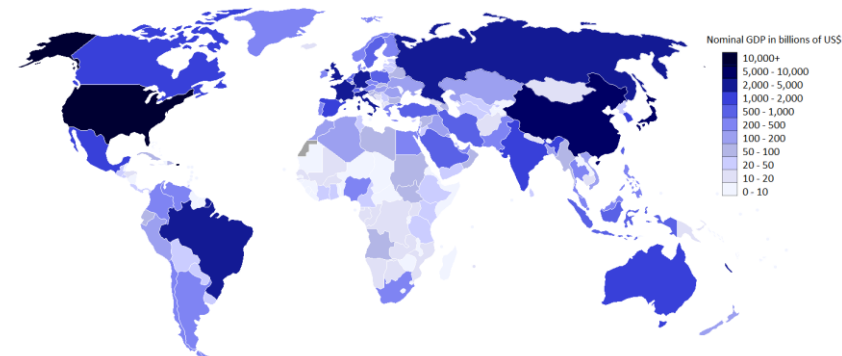


As formas da distribuição permitem interpretações efetivas sobre os fenômenos e permite modelagem destas por funções que facilitam as análises

Valores realmente significativos

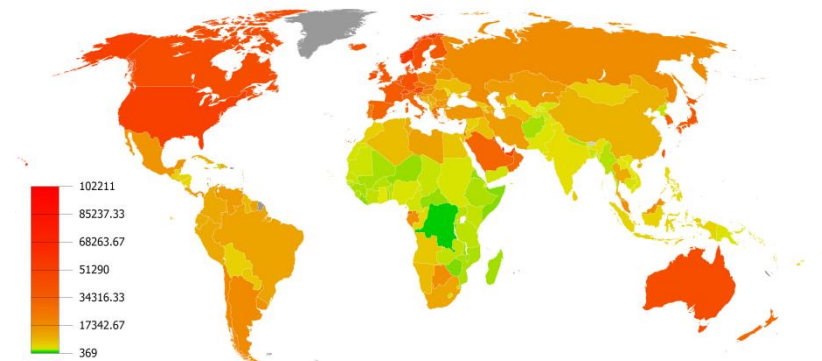
- Absolutos ou relativos?
 - Qual é o valor realmente significativo?
 - Qual é o valor que realmente vale a pena mostrar?

PIB



- Exemplos
 - PIB x Renda per capita
 - Potência x Rendimento
 - No. de estudantes x No. de estudantes por docente

Renda per capita



Tamanho x Qualidade

Parâmetros adimensionais

Os parâmetros mais importantes de um fenômeno em estudo



A relação entre eles

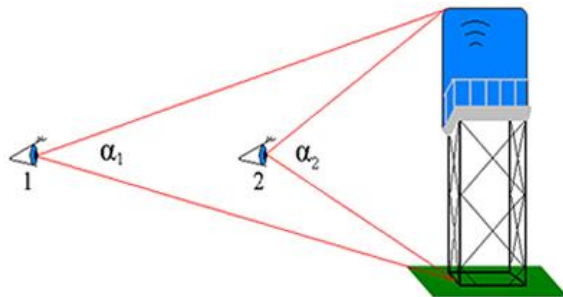
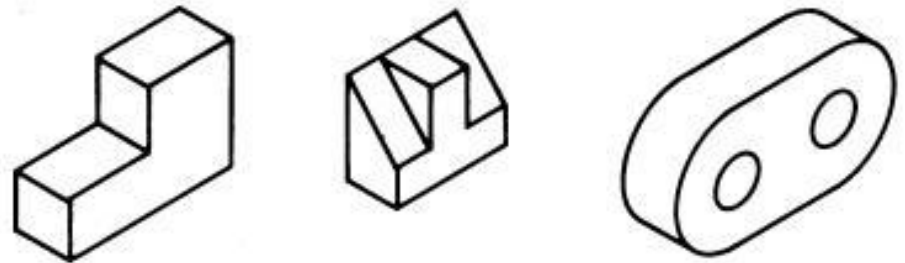
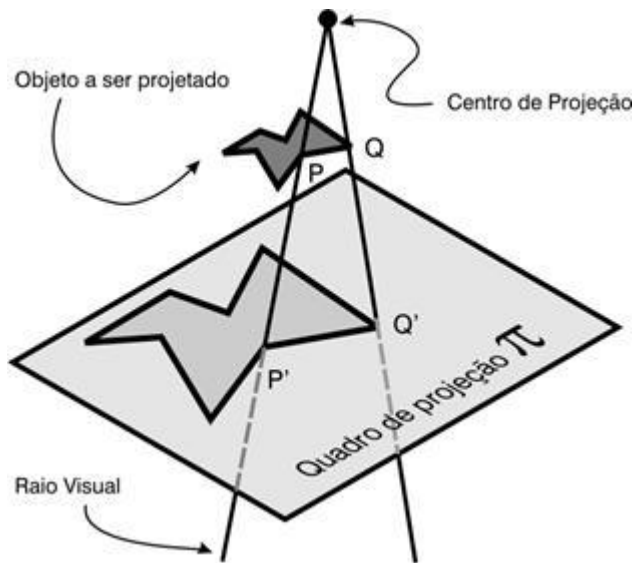


Permite servir de referência para:

- definir regimes diferentes do fenômeno
- identificar situações semelhantes e
- generalizar os resultados!

Parâmetros adimensionais

- Exemplo: visão cônica e visão cilíndrica



Relação determinante: D/d !!!

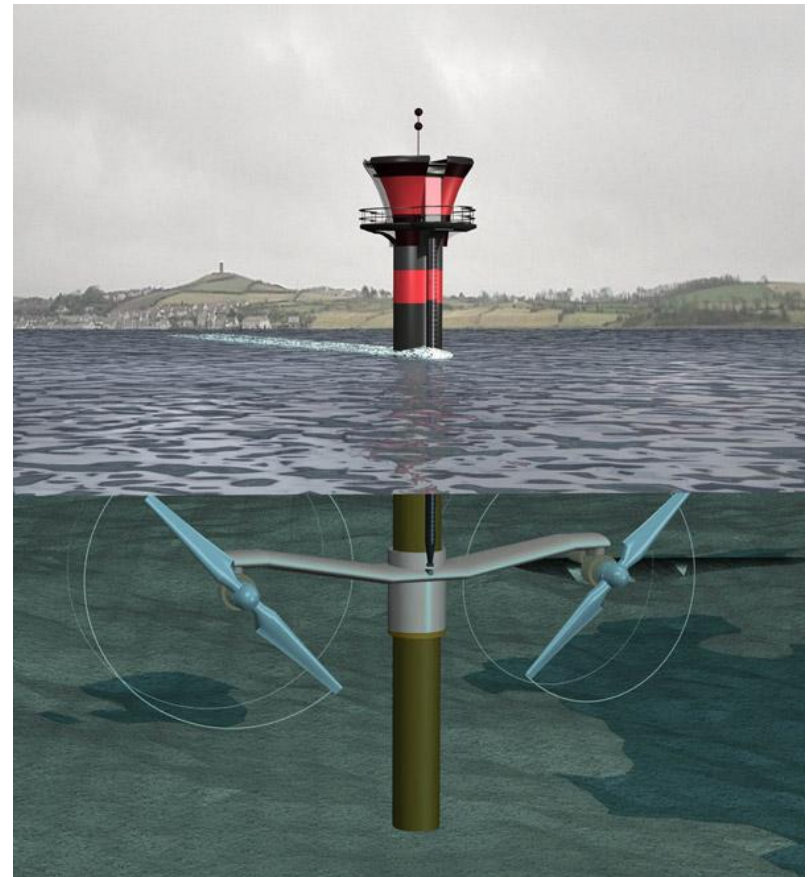
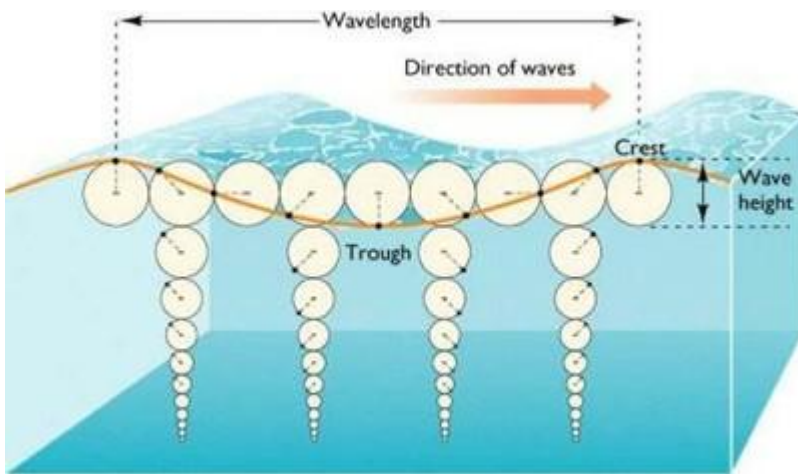
D: tamanho do objeto;

d: distância

Parâmetros adimensionais

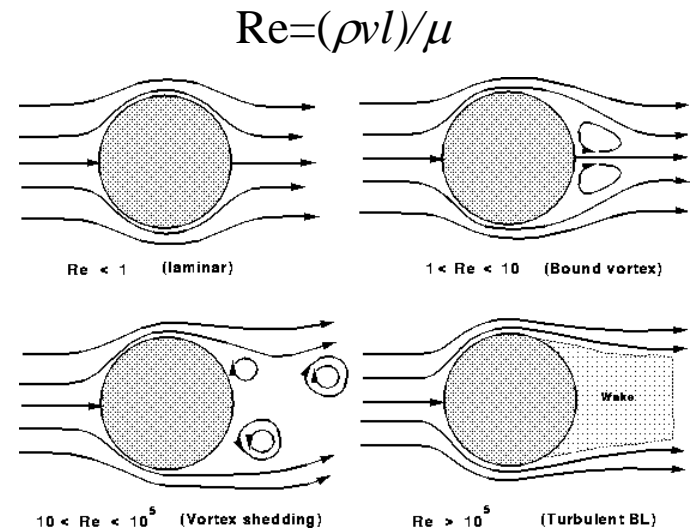
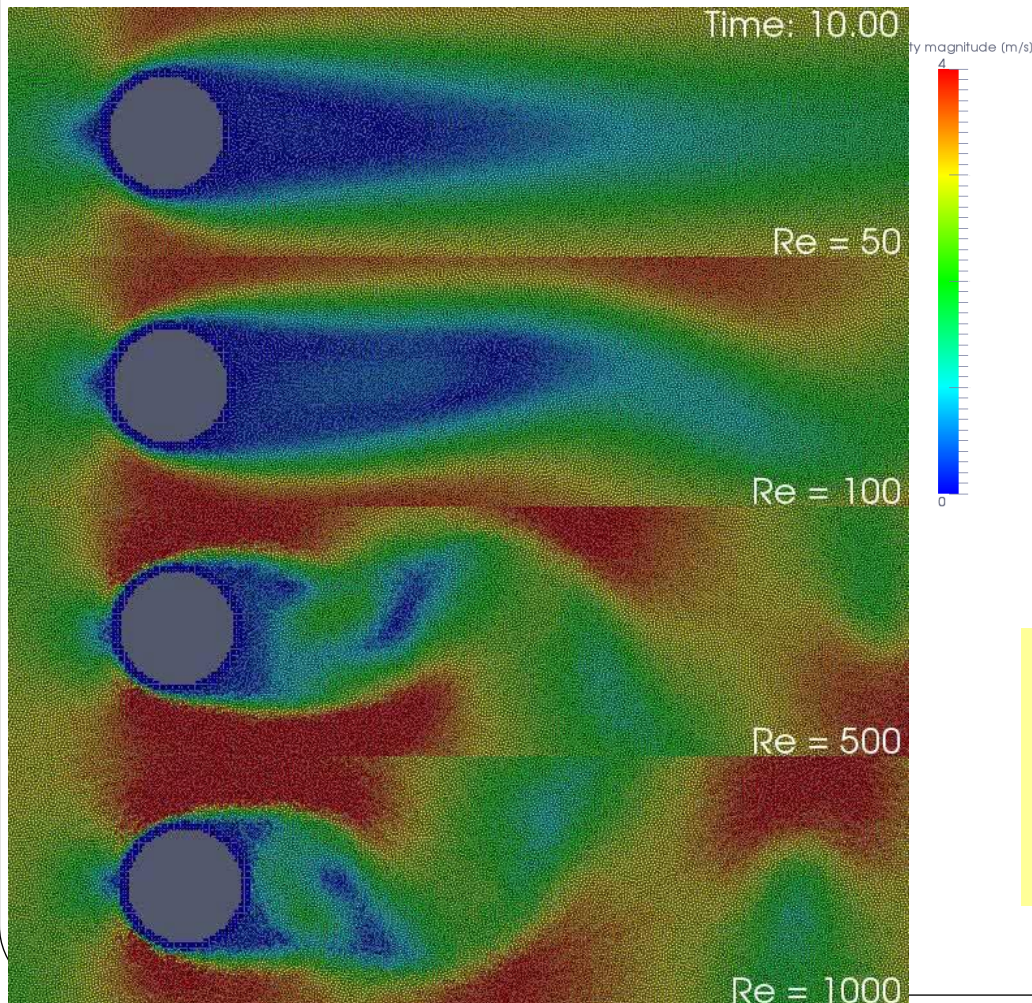
- Exemplo: Coluna vertical sob ação de ondas

Diferentes regimes de interação entre a coluna e as ondas classificados usando o parâmetro D/λ



Números adimensionais

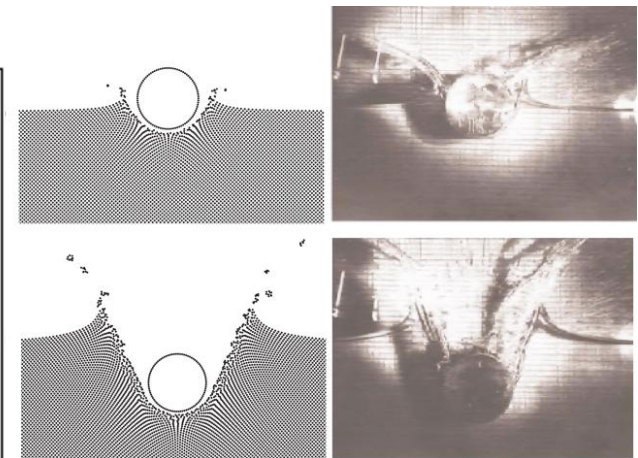
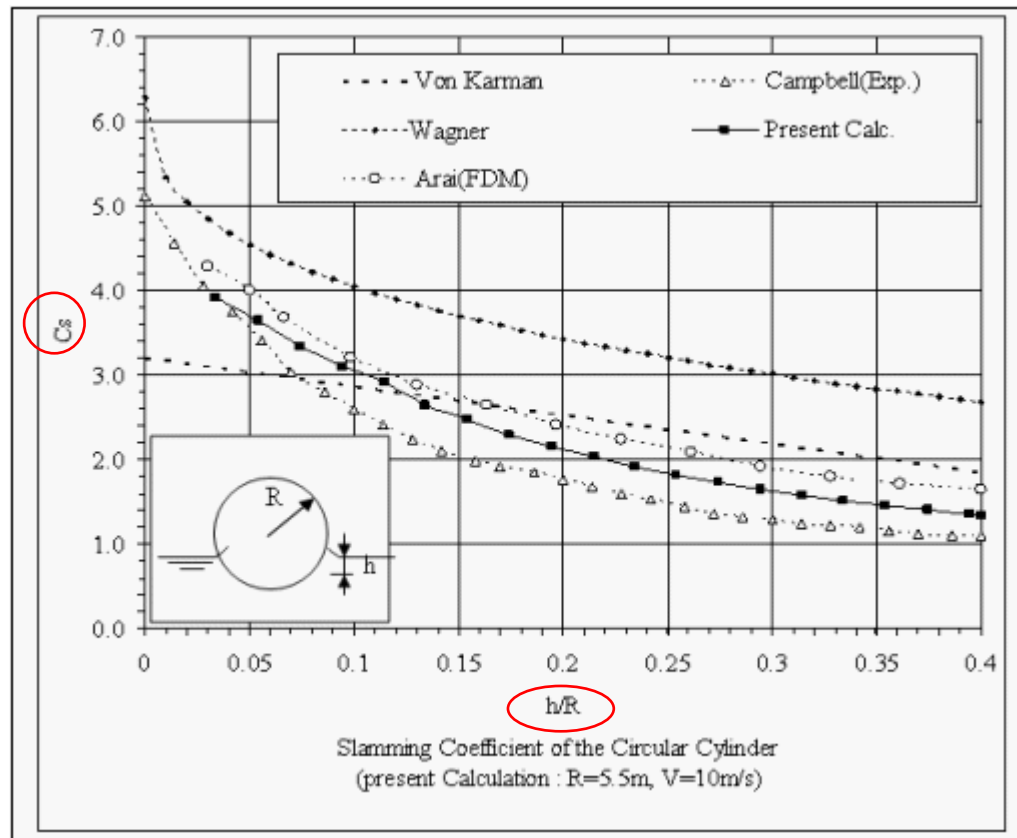
- Exemplo: Esteira de um cilindro



Não importa qual a dimensão, velocidade, densidade ou viscosidade, o comportamento dinâmico depende exclusivamente do adimensional $Re!!!$

Porque isso é importante?

- Resultados com elevada densidade de informação
- Exemplo: Força de impacto durante a queda de um cilindro na água.



Usando os adimensionais que são realmente significativos, aumenta a densidade da informação e permite generalizar a conclusão.

Epílogo

- Filme da semana: Interstellar

“...explorers travel through a wormhole in space...”

