AULA 1

Introdução à visualização de dados

O que é a visualização de dados?

- É focado em como representar visualmente e explorar uma grande quantidade de dados.
- Tirar vantagens das capacidades visuais humanas
- Fornecer "insights" sobre o fenómeno por detrás dos dados

Definição

A representação e apresentação de dados que exploram o nosso visual habilidades de perceção para amplificar a cognição.

Benefícios da visualização

- Ajuda a pensar
- Reduz a carga na memória de trabalho
- Descarrega a cognição
- Usa o poder da perceção humana

Dados e visualização de informação

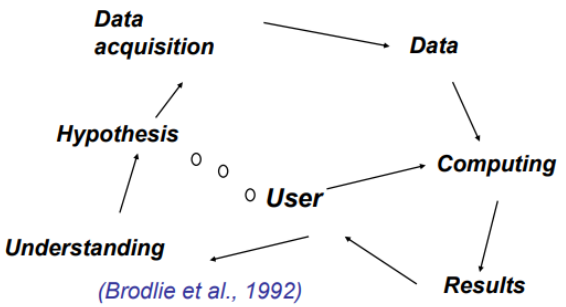
- Visualização de dados – científicos (DV) – Os dados têm uma estrutura espacial inerente.

Visualização de informação (IV) - Dados tabulares "abstratos" sem estrutura espacial inerente (dados tabulares)

- Estas designações podem ser errôneas; ambos começam com dados (crus) e permitem a extração de informação
- Os limites entre ambas as áreas não é bem explicito.

Estrutura

A visualização não inclui apenas a construção de imagens/gráficos a partir de dados, mas também as suas transformações e manipulações (se possível a sua aquisição):



É um problema "**human-in-the-loop**"

Aplicações

Visualização científica:

- Todos os campos da engenharia;
- Medicina;
- Metereologia, climatologia, oceanografia;
- …

O problema "human-in-the-loop" envolve o utilizador como uma parte do sistema

São bastante complexos devido ao facto de:

- Os humanos serem sistemas muito complexos
- Não se sabe bem
- No geral não podemos mudá-los

O perfil, as necessidades e o contexto de uso dos usuários-alvo deve ser cuidadosamente considerado ao projetar uma visualização.

Os dados podem ser

- Simulados
- Medidos a partir de um fenómeno real

Depois é aplicada uma técnica de visualização envolvendo:

- **Transformando** os dados através de alguns métodos
- **Mapeando** uma forma adequada de representar os dados visualmente
- Produzir uma imagem ou uma sequência de imagens (**renderizar**)

Este processo é repetido as vezes necessárias até trazer um **insight**

A escolha do mapeamento mais correto é fundamental!

Considerando os valores de uma função ou os dados relativamente à altitude de um terreno, ou a profundidade do mar:

Podem ser usadas **diferentes técnicas de visualização e mapeamento**:

- Superfície 3D
- Pseudo-cores
- Contornos (isolinhas)

Independentemente do propósito a visualização

- Deve permitir descarregar a cognição interna e o uso da memória para o sistema percetivo, usando imagens cuidadosamente projetadas como uma forma de representações externas (memoria externa)
- Para apoiar as tarefas dos usuários

Para criar visualizações simples ou complexas é importante encontrar quais as questões fundamentais que o usuário irá fazer.

AULA 2

Dados

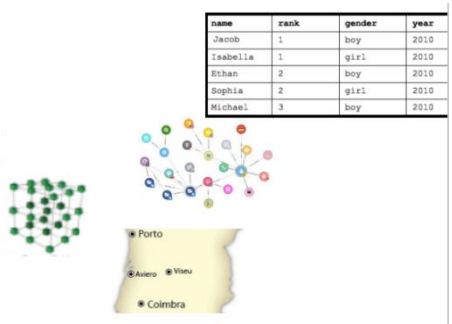
Características dos dados

A classificação de dados é importante para:

- Prever qual visualização é a mais adequada
- Ser mais fácil de comunicar sobre os dados
- Permite uma abordagem mais sistemática da visualização

Abstração de dados

- Quatro tipos básicos de conjuntos de dados:
  Tabelas
  Redes
  Geometria
  Campos



- Cinco tipos de dados básicos:
  Itens
  Atributos
    Categóricos
    Ordenados
      Ordinais
      Quantitativos
  Ligação
  Posições
  Grades

Níveis de representação de dados

- Qualitativo (ou categórico)
- Quantitativo (ou numérico)

Natureza dos dados

- Continua
- Discreta

Escala de medida

- Nominal
- Ordinal
- Intervalo
- Razão

EXEMPLOS DE ESCALAS DE MEDIDA

NOMINAL – marcas de carro, género, espécies de animais, ...

ORDINAL – dias da semana, preferências …

INTERVALO – data, IQ, temperaturas em ºC

RAZÃO – temperaturas em K, peso, altura, …

- A escala de razão **nível mais alto de representação**, tem um não zero arbitrário (contrariamente à escala de intervalo)
- Às vezes a estrutura dos dados não é suficiente
- É necessário saber o fenómeno por detrás dos dados tal como saber as necessidades (questões) dos utilizadores.

Preparação dos dados

Algumas fases e termos:

- Data pre-processing
- Data wrangling
- Data cleaning, Data tiding …
- Data transformation

A integridade dos dados torna-se mais essencial quando o volume dos dados aumenta

Limpeza de dados

- Codificação
- Agregação
- Dados derivados
- Remoção
- Normalização

Examples:

Cleansing Data

Birth date: Feb/30/2000
Temperature: -300 °K
City: Lixboa

Transforming Data

- Encoding – answers to an open question need to be parsed and coded
- Aggregation – detail may be excessive (age: <18; 18-40; 41-65; >65)
- Derived data – add new relevant variables ($T_{range} = T_{max} - T_{min}$)
- Removal – remove data that are not needed
- Standardization – M/F; °C or °F

AULA 3

Criar uma visualização

Mapeamento visual

É necessário decidir:

- Qual a estrutura de visualização usada para representar os dados
- A sua localização no display

Alguns tipos de dados podem ser facilmente mapeados por uma localização espacial.

Dados abstratos não têm uma fácil correspondência com as dimensões do espaço físico à nossa volta.

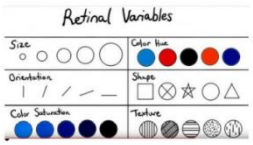
Três estruturas podem definir o mapeamento visual/ codificação:

- **Substrato espacial** – dimensões no espaço físico onde a representação visual é criada (pode ser definida em termos de eixos e tipos de dados)
- **Elementos gráficos** – qualquer coisa visível aparece no espaço (pontos, linhas, superfícies, volumes)
- Propriedades gráficas – propriedades dos elementos gráficos para os quais a retina humana é sensível – **Variáveis de retina**:
- Tamanho, orientação, cor, textura e forma

- **Spatial substrate** axes (x, y, …)
  type of data (quantitative, ordinal, categorical)

- **Graphical elements** points
  lines
  surfaces
  volumes

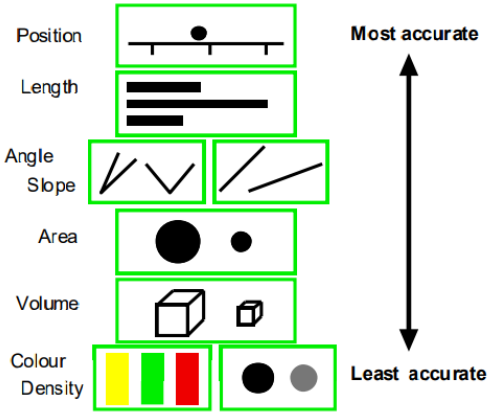
- **Graphical properties** retinal variables:
  size,
  orientation
  color (depends on physiology and culture)
  texture
  shape



Como selecionar codificação visual?

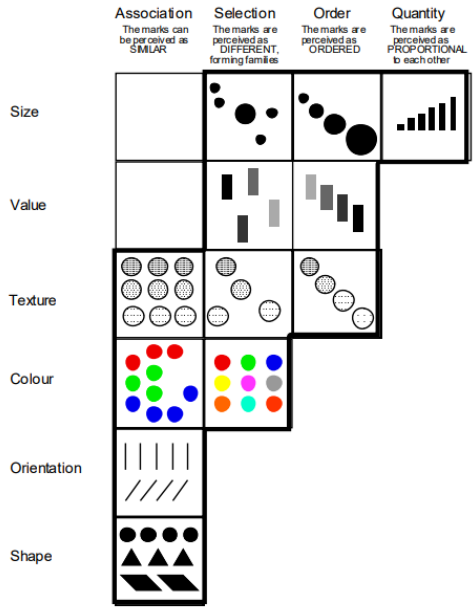
Interpretação da orientação de Bertin sobre a adequação de vários métodos de codificação para suportar tarefas comuns.

Note que apenas o tamanho é adequado para representar a quantidade com precisão.



Como escolher a descodificação visual para representar com precisão uma quantidade?

A relativa dificuldade de avaliar o valor quantitativo em função do mecanismo de codificação, conforme estabelecido por Cleveland e McGill



Mapeamento – valor visual de codificação

Alguns atributos visuais como o tamanho, proximidade são também rapidamente processados pela precessão visual, mesmo antes do processo cognitivo.

Procedimentos a seguir para criar representações visuais de dados

1. Definir o problema e as questões do utilizador

2. Examinar a natureza dos dados para representar o pré-processamento dos dados
3. Definir o número dos atributos/ variáveis / dimensões
4. Escolher a estrutura visual para mapear.

Importantes aspetos a considerar:

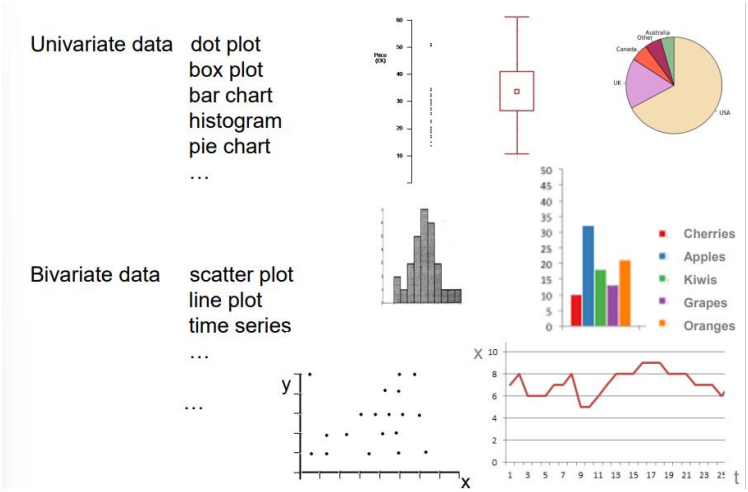
nature of the problem — communicate, explore, confirm

nature of the data to represent — quantitative, ordinal, categorical

number of attributes — univariate, bivariate, trivariate, multivariate

dataset types — tables, networks, spatial or geographical, fields, geometry

Técnicas de visualização comuns para representar visualmente dados univariados e bivariados



Representação de dados univariados

Diferença entre histograma e gráfico de barras:
Histograma – representa uma distribuição numérica

Gráfico de barras – representa o número de ocorrências de dados categóricos ou ordinais

Gráfico circular – representa uma proporção numeral, parte de um inteiro. Alguns especialistas recomendam que estes gráficos sejam evitados uma vez que é difícil comparar secções do gráfico ou dados entre diferentes gráficos circulares.

Quais os critérios necessários para criar um gráfico circular:

- As partes fazem um todo significativo
- As partes são mutualmente exclusivas
- Existem <6 partes e as fatias não têm tamanhos muito diferentes

Se o propósito é comparar pates, use outro gráfico!

Representação de dados bivariados

O scatterplot é a representação mais convencional

Cada observação é representada por um ponto num espaço bidimensional.

Os eixos são associados a dois atributos

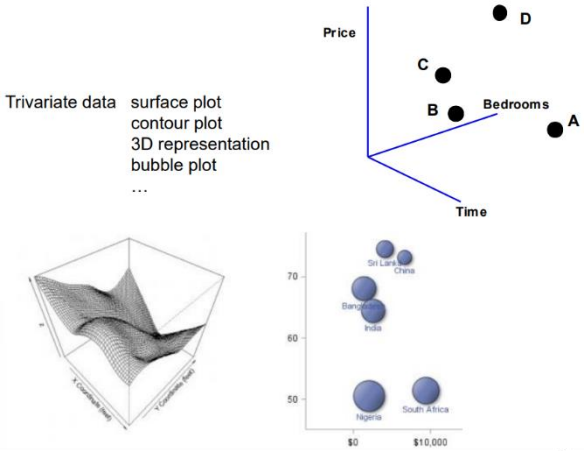
Esta representação requer consciência para:

- Tendências locais
- Outliers
- Trocas locais

Gráficos de linha, plot lineares ou gráficos de curva providenciam informação como uma serie de pontos chamados “marcadores” conectados por segmentos de linha retos.

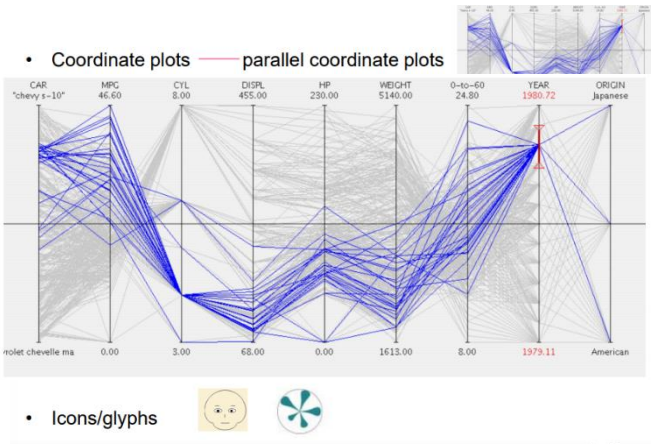
Representação de dados trivariados

A representação 3D pode ser ambígua, para que ela nos seja mais útil deve ser capaz de ser alterada.



Num gráfico de bolhas os dados são representados como discos que expressão dois dos valores através da localização dos discos no plano xy e o terceiro através do seu tamanho.

Representação de dados multivariados (ou hipervariados)



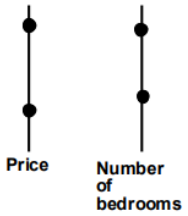
Parallel coordinates plots é uma das técnicas mais populares para os dados hipervariados

Como fazer estas gráficos?
Considerando um caso simples de dado bivariáveis

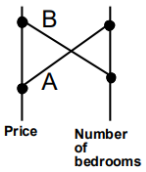
EXEMPLO:

1. Um scatterplot representa o preço e o número de quartos associados com duas casas

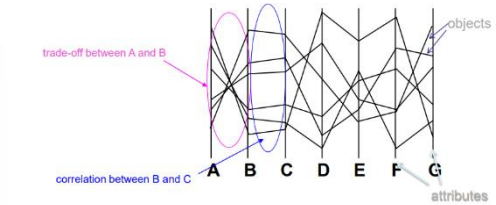
2. Os eixos são destacados e paralelos; cada casa é representada por um ponto em cada eixo



3. Para evitar ambiguidade nos pares de pontos que representam a casa são juntos e legendados.

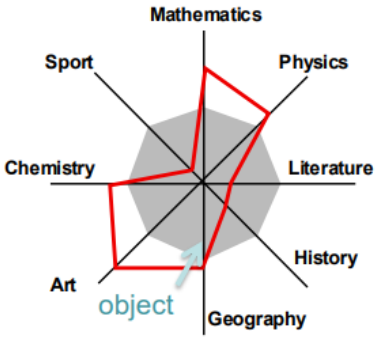


Para objetos caracterizados por muitos atributos o parallel coodinate plots oferecem muitas vantagens.



The trade-off between A and B, and the correlation between B and C, are immediately apparent. The trade-off between B and E, and the correlation between C and G, are not.

Star plots tem muitos recursos em comum com o parallel coordinate plots.

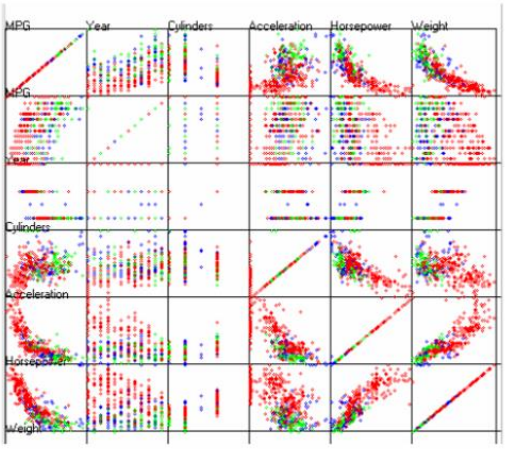


- Um valor atribuído é representado por um ponto num eixo de coordenadas
- Os eixos atribuídos irradiam desde uma região comum
- Para um dado objeto, os pontos são juntos por linhas retas

Propriedades

- As suas formas permitem uma apreciação rápida relativamente aos atributos dos objetos
- Oferecem visibilidade dos objetos e são adequados para comparar com outros

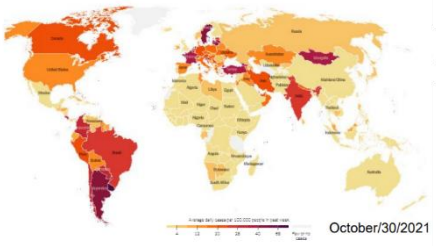
O



Scatterplot matrix for 6 attributes of a car dataset

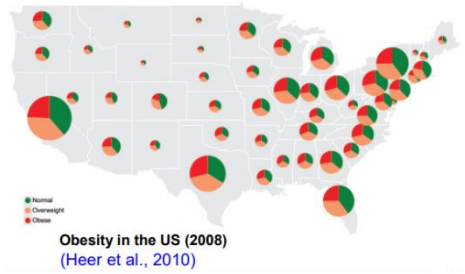
scatterplot mtrix (SPLOM) é aplicável para um grande número de variáveis.

Mapas coropléticos – Uma abordagem padrão para comunicar dados agregados por áreas geográficas usando a codificação de cores em área geográfica



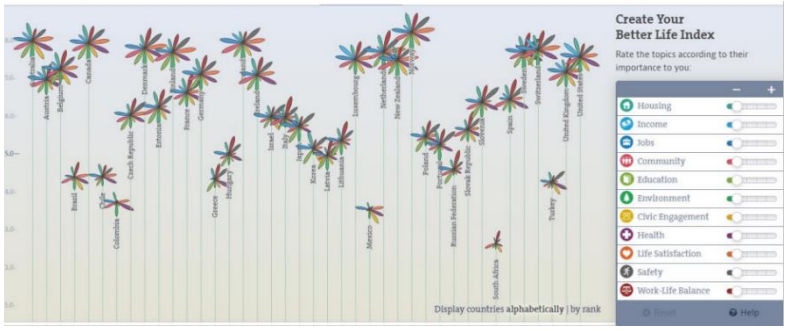
Graduated symbol maps são uma alternativa aos mapas coropleticos.

- Os símbolos são colocados sobre um mapa subjacente; pode mostrar mais dimensões
- Evita confundir área geográfica com valores de dados



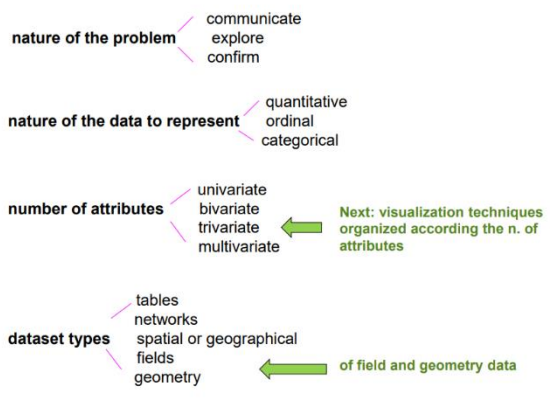
Exemplo de gráfico de glifos

As propriedades físicas da forma representam diferentes variáveis categóricas dimensionando de acordo com o valor quantitativo associado e distinguido pela cor



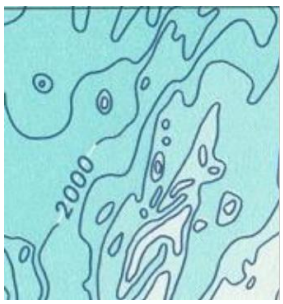
Dashboards

Display visual que sumariza um conjunto de dados oferecendo informação mais rápida.



Representação de um escalar num campo 2D

Gráficos de contorno

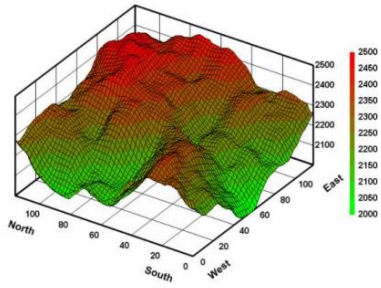


- Linhas de contorno (também isolinhas, isopleth ou curva equipotencial) de uma função de duas variáveis é uma curva ao longo da qual a função tem um valor constante, de modo que a curva una os pontos de igual valor.
- Típico para gráficos de metrologia (curvas isobáricas e isotérmicas) e mapas (para representar altitudes ou profundidades)

Representação de uma função escalar em 3D

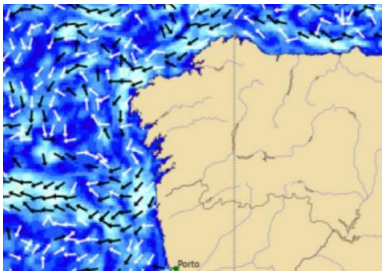
Gráficos de superfície

- Podem ser combinados com cores (preferencialmente numa forma redundante e escolhendo cuidadosamente a escala)



Representação de vetores num campo 2D

- Representam principalmente correntes do mar à superfície



Representação de vetores num campo 3D

Glifos – four dimensional data visualization

AULA 4
Visualização efetiva

Implica dizer a **verdade** relativamente aos dados

Há métodos para avaliar visualizações que devem ser usados aquando do processo de criação de uma visualização.

A precessão varia de pessoa para pessoa dependendo do contexto e da experiência da mesma.

O que fazer quando queremos representar dados em gráficos?

Dois fundamentos principais:

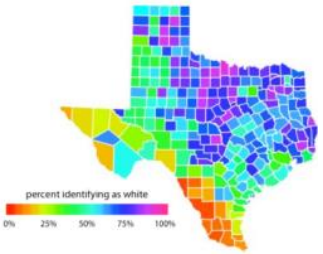
- Representar números como uma medida física na superfície do gráfico, **diretamente proporcional** à quantidade numérica representada.
- **Rotulação** clara e completa para eliminar qualquer distorção gráfica ou ambiguidade

NOTA: A representação visual deve ser **testada** quanto à sua eficiência e eficácia para os usuários alvo e relativamente às duas tarefas.

As cores podem usar bastante, no entanto podem atrapalhar quando mal aplicadas… devem ser usadas com moderação e cuidado, deve servir a um propósito, ser claro e não distrair.

Armadilhas comuns no uso de cores

- **Codificar informações demais ou irrelevantes** (no exemplo dos estados a coloração não serve para nada, as cores são demasiado saturadas e são uma distração)



- **Usar escalas de cores não apenas num tom para codificar valores de dados** (a escala de cores do arco-íris não é muito apropriada para visualizar valores de dados contínuos; tende a enfatizar características arbitrarias dos dados)
- **Não projetar para deficiência de visão de cores**

Guidelines para usar cores

- Design primeiro em preto e branco; depois aplique a cor com moderação e cuidado
- Use rotulagem direta em vez de cores quando precisar distinguir entre mais de cerca de oito itens categóricos
- Evite grandes áreas preenchidas de cores excessivamente saturadas. Eles dificultam para o leitor inspecionar cuidadosamente a figura
- Para garantir que os números funcionem para pessoas com cvd, não confie apenas em escalas de cores específicas. Em vez disso, teste figuras em um simulador de cvd
- Não use azul para colorir objetos pequenos (será difícil de ver)