GENE7033 – Tópicos Especiais em Genética I:

Visualização de dados para publicações científicas

Profª Drª Chirlei Glienke

Drª Desirrê Petters-Vandresen

Tipos de dados, variáveis e elementos estéticos

Drª Desirrê Petters-Vandresen

13/10/2022

Convertendo dados em elementos estéticos

- Especialmente no meio acadêmico, visualização de dados é uma das habilidades mais importantes (principalmente pela elaboração de figuras/gráficos)

- Figuras reforçam os argumentos, precisam ser claras, bem elaboradas e convincentes, e podem fazer a diferença entre:
  - Um artigo que é aceito ou não
  - Um artigo que se torna muito citado e influente em uma área e um artigo que é esquecido

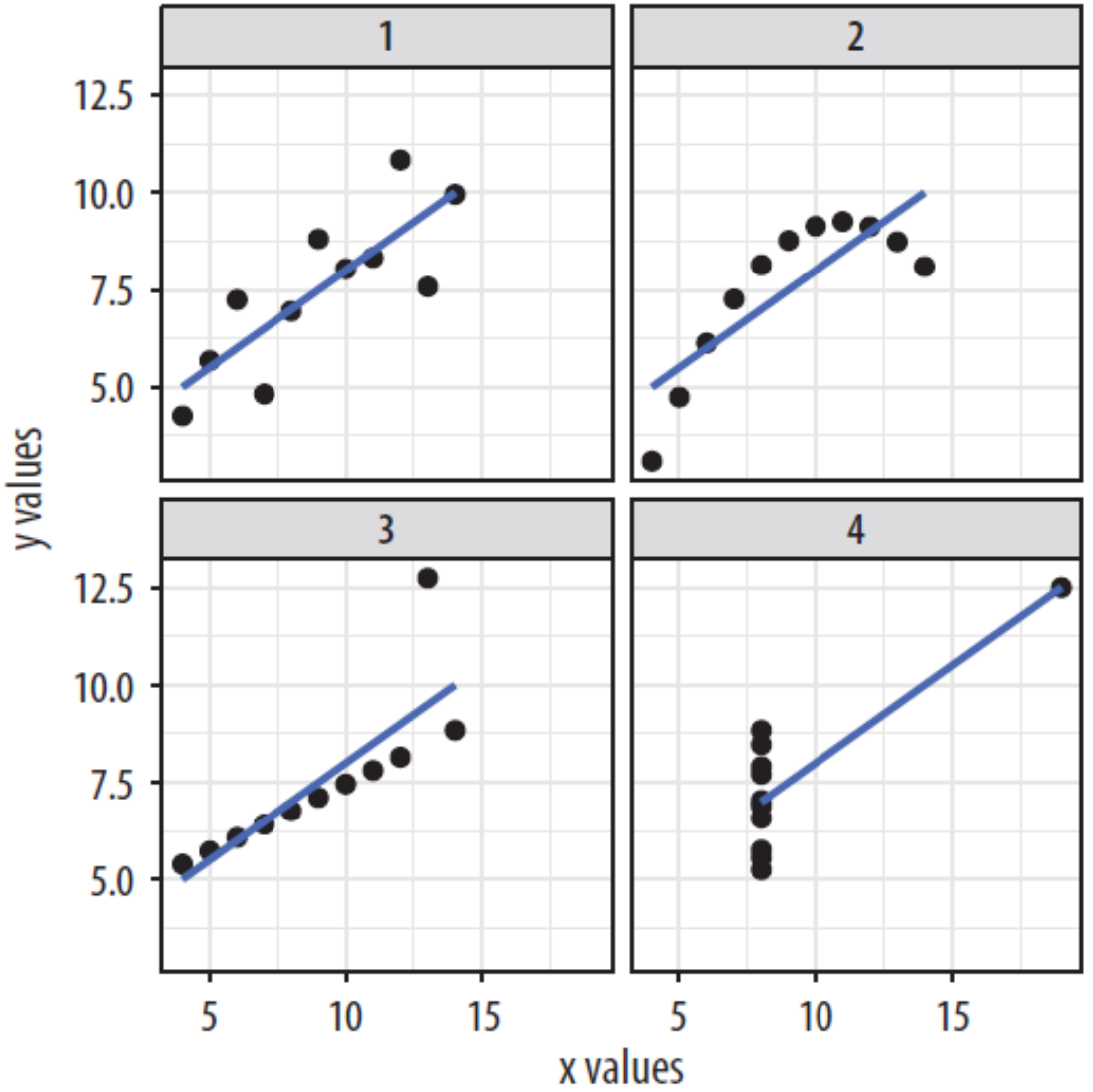
Por que visualizar dados?

"*Never trust summary statistics alone; always visualize your data*"
(Alberto Cairo)

"*Few of us scape being indoctrinated with these notions: (1) numerical calculations are exact, but graphs are rough (...)*" (Francis Anscombe)

Por que visualizar dados?

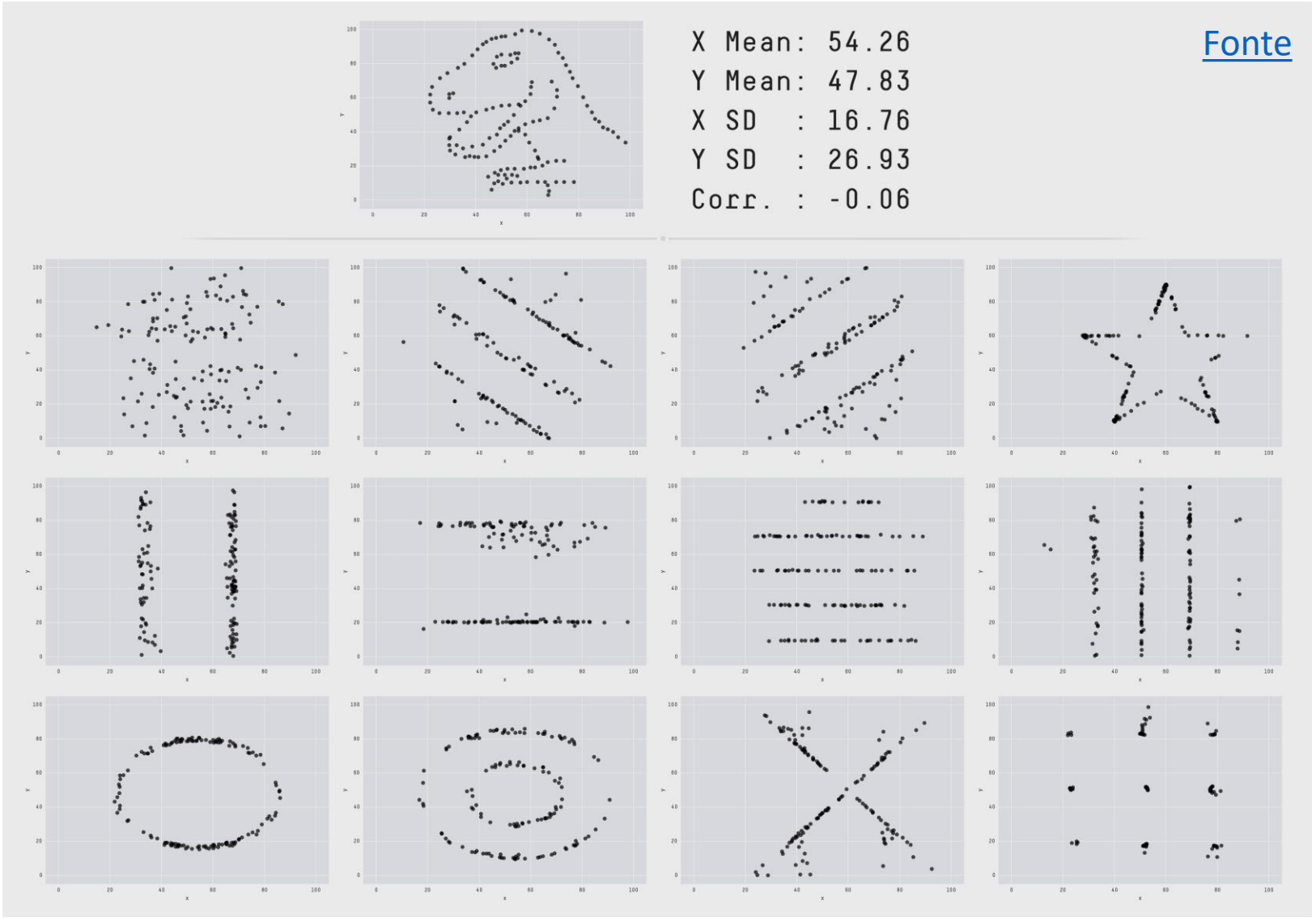
- Quarteto de Anscombe (1973)
  - Gráficos de dispersão (relação entre duas variáveis *x* e *y*)
  - Onze observações para cada variável
  - Propriedades numéricas das variáveis (média, variância, correlação entre variáveis) idênticas para os quatro conjuntos



Adaptado de:
HEALY, K. **Data Visualization: A Practical Introduction.** Princeton University Press, 2019.

Por que visualizar dados?

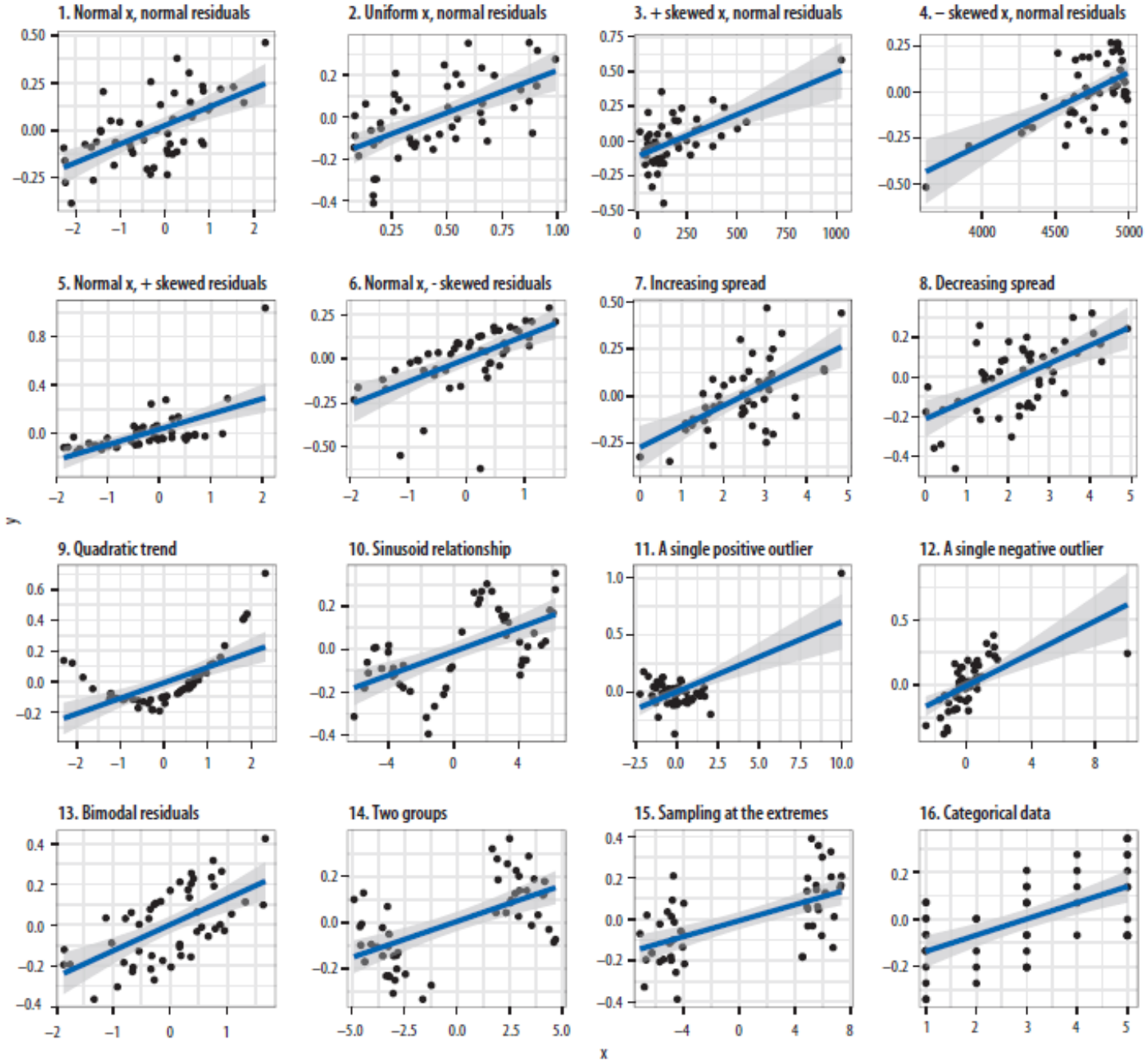
- Datasaurus Dozen (Matejka & Fitzmaurice / Cairo)
  - Gráficos de dispersão (relação entre duas variáveis *x* e *y*)
  - Propriedades numéricas das variáveis (média, variância, correlação entre variáveis) idênticas para os conjuntos



Fonte

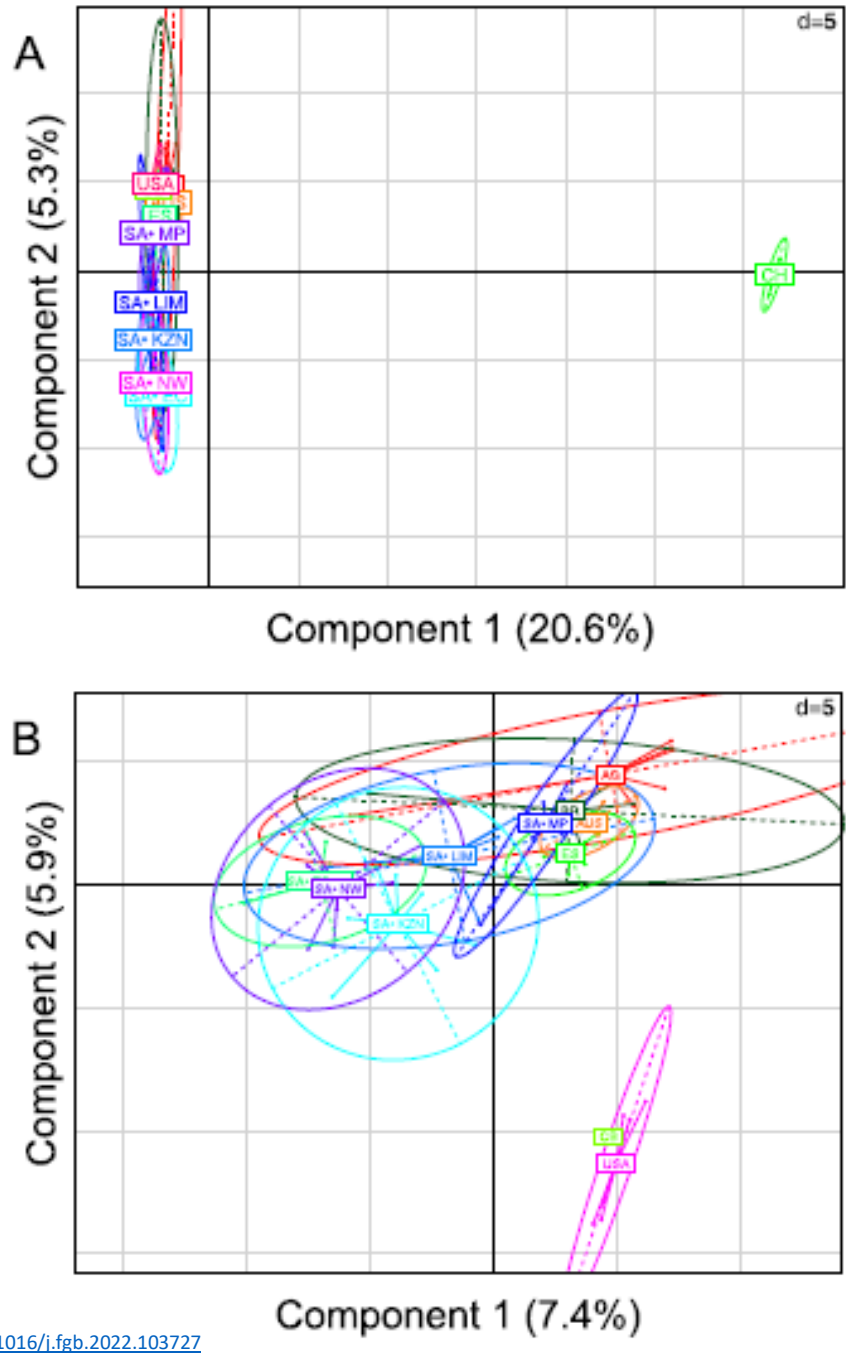
Por que visualizar dados?

- Jan Vanjove – Quais padrões podem estar por trás de uma correlação?
  - Para todos os gráficos, o coeficiente de correlação é de 0.6
  - Para mais detalhes:
  - https://janhove.github.io/teaching/2016/11/21/what-correlations-look-like



Por que visualizar dados?

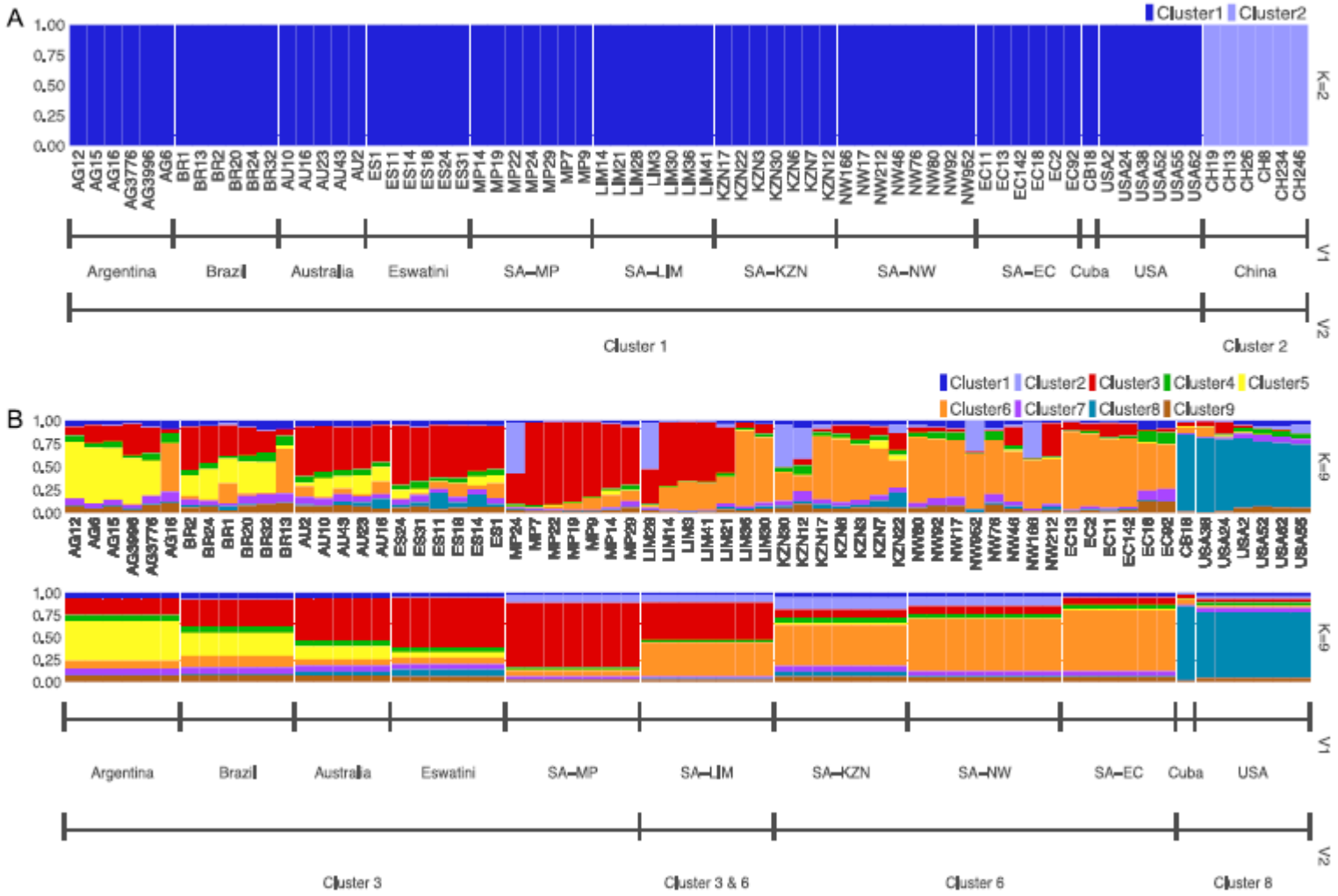
- Análise de componentes principais
  - Investigação da estrutura populacional do fitopatógeno de citros *Phyllosticta citricarpa* realizada por Coetzee e colaboradores
  - A inclusão de dados de SNP de uma população geneticamente distante (China) diminui a resolução e visualização de diferenças entre as outras populações



Adaptado de:
COETZEE et al. 2022. **Fungal Genetics and Biology**. DOI: 10.1016/j.fgb.2022.103727

Por que visualizar dados?

- Análise de componentes principais
  - Investigação da estrutura populacional do fitopatógeno de citros *Phyllosticta citricarpa* realizada por Coetzee e colaboradores
  - A inclusão de dados de SNP de uma população geneticamente distante (China) diminui a resolução e visualização de diferenças entre as outras populações



Adaptado de:
COETZEE et al. 2022. **Fungal Genetics and Biology**. DOI: 10.1016/j.fgb.2022.103727

Introdução

- Especialmente no meio acadêmico, visualização de dados é uma das habilidades mais importantes (principalmente pela elaboração de figuras/gráficos)

- Figuras reforçam os argumentos, precisam ser claras, bem elaboradas e convincentes, e podem fazer a diferença entre:
  - Um artigo que é aceito ou não
  - Um artigo que se torna muito citado e influente em uma área e um artigo que é esquecido

Introdução

- Escassez de recursos para ensino e desenvolvimento de habilidades em visualização de dados
  - Poucas disciplinas na área
  - Literatura não tão abundante
  - Tutoriais normalmente focados em como atingir certos efeitos visuais (ex: troca de cor e fonte, tamanho de elementos do gráfico) e não nos motivos pelos quais certos tipos de gráficos são melhores que outros em determinadas situações

Situação semelhante ao processo de desenvolvimento das habilidades de escrita acadêmica

“Ouvido” e “olhar”

- “Ouvido” para a escrita: habilidade de ouvir (internamente) quando a escrita é boa

- “Olhar” para a elaboração de figuras: habilidade de observar uma figura e perceber se está clara, convincente, atrativa e balanceada

- Assim como o “ouvido”, o “olhar” pode ser treinado, aprendendo algumas regras e princípios gerais, e passando a prestar atenção em detalhes que muitas vezes passam despercebidos

Para desenvolver o “olhar”

- Assim como o “ouvido”, não se treina o “olhar” em um fim de semana, lendo um livro ou assistindo a somente um curso
- Muitos conceitos são complexos e precisam de tempo para serem assimilados
- Exposição à novas abordagens, olhar atento ao processo alheio de elaboração de figuras e escolhas metodológicas
- Mente aberta para mudar de ideia e disposição para tentar várias e várias vezes até estar satisfeito com uma figura

"Data visualization is part art and part science. The challenge is to get the art right without getting the science wrong and vice versa."
(C. Wilke)

- 1. As técnicas de visualização de dados devem representar os dados de forma adequada, sem distorções ou margens para erros de interpretação

- 2. Representar os dados de forma esteticamente agradável e adequada é essencial para reforçar a mensagem sendo transmitida

- Cores desarmônicas, elementos desbalanceados distraem o leitor, dificultando a interpretação e entendimento da figura

Em geral...

- Cientistas/pesquisadores:
  - Hábeis em visualizar os dados sem distorcer a informação ou gerar erros de interpretação
  - Senso estético pouco desenvolvido, dificuldade em escolher os elementos visuais adequados para reforçar a mensagem

- Designers:
  - Hábeis em preparar figuras esteticamente agradáveis e bonitas
  - Pouca atenção aos dados em si, gerando distorções na informação e erros

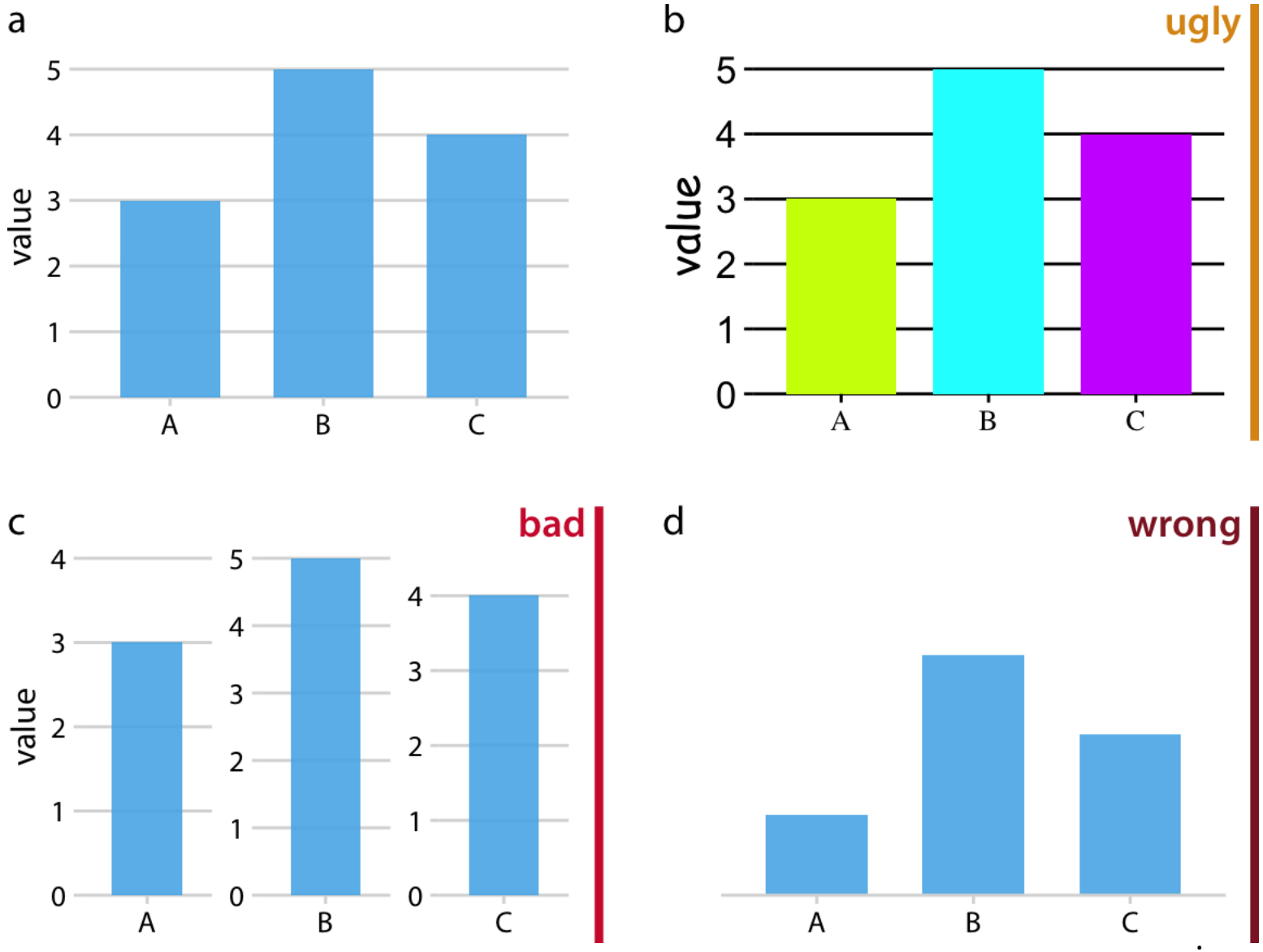
Como obter o “melhor dos dois mundos?”

Para nossa “sorte”...

- Não é somente uma questão de “bom-gosto”
- Não é somente uma questão de opinião pessoal e subjetiva sobre o que é bonito ou feio
- Trata-se mais de uma questão sobre como nossa percepção visual funciona, e como nosso cérebro percebe diferentes elementos visuais como comprimento, tamanho absoluto e relativo, orientação, formas e cores
- Aprender sobre estes aspectos e elementos visuais é mais fácil do que simplesmente “*desenvolver o feeling*” para o que é uma boa figura

O que constitui uma figura ruim?

- Normalmente uma mistura de “mau-gosto”, problemas com os dados ou problemas de percepção em função do tipo de gráfico escolhido
- Figuras feias: problemas estéticos, porém clara e informativa
- Figuras ruins: problemas relacionados à percepção, figura confusa, complicada ou que gera erros de interpretação
- Figuras erradas: problemas matemáticos, objetivamente incorreta (do ponto de vista dos dados)



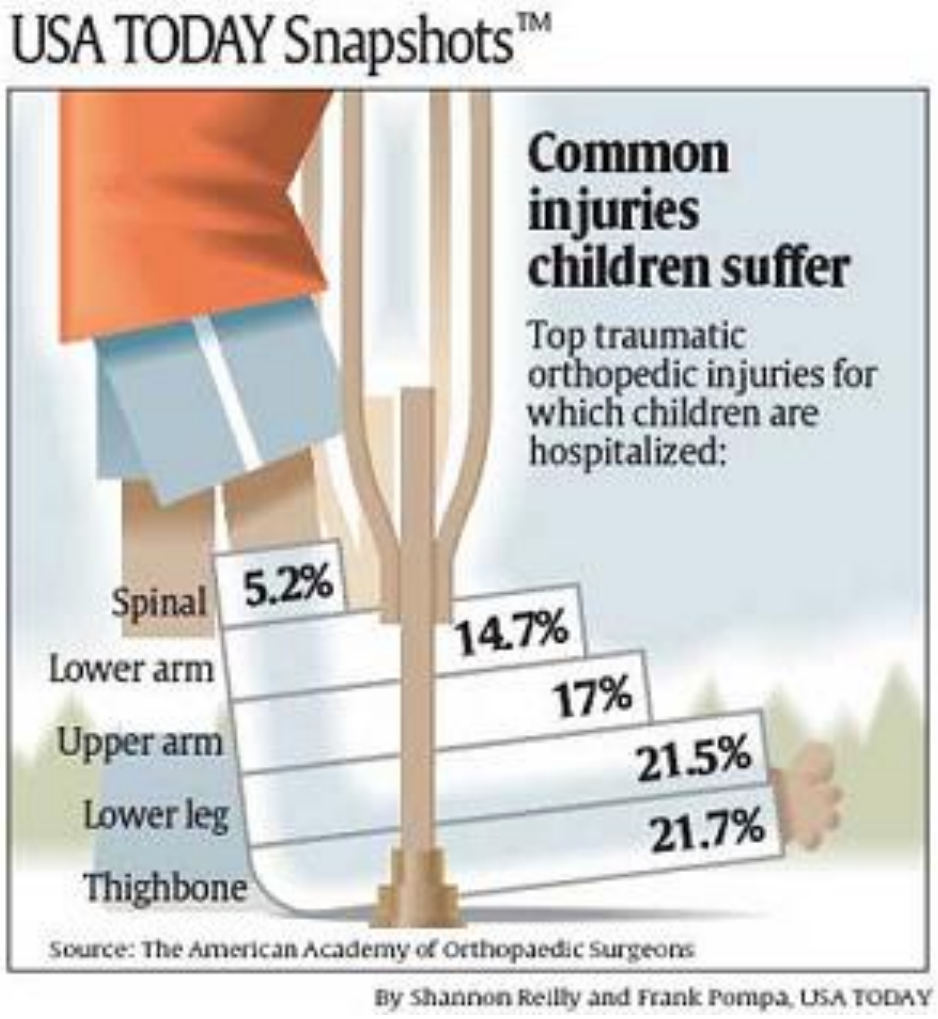
Adaptado de:
WILKE, C. O. **Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures.** O’Reilly Media, 2019.

“Mau-gosto” ou problemas estéticos

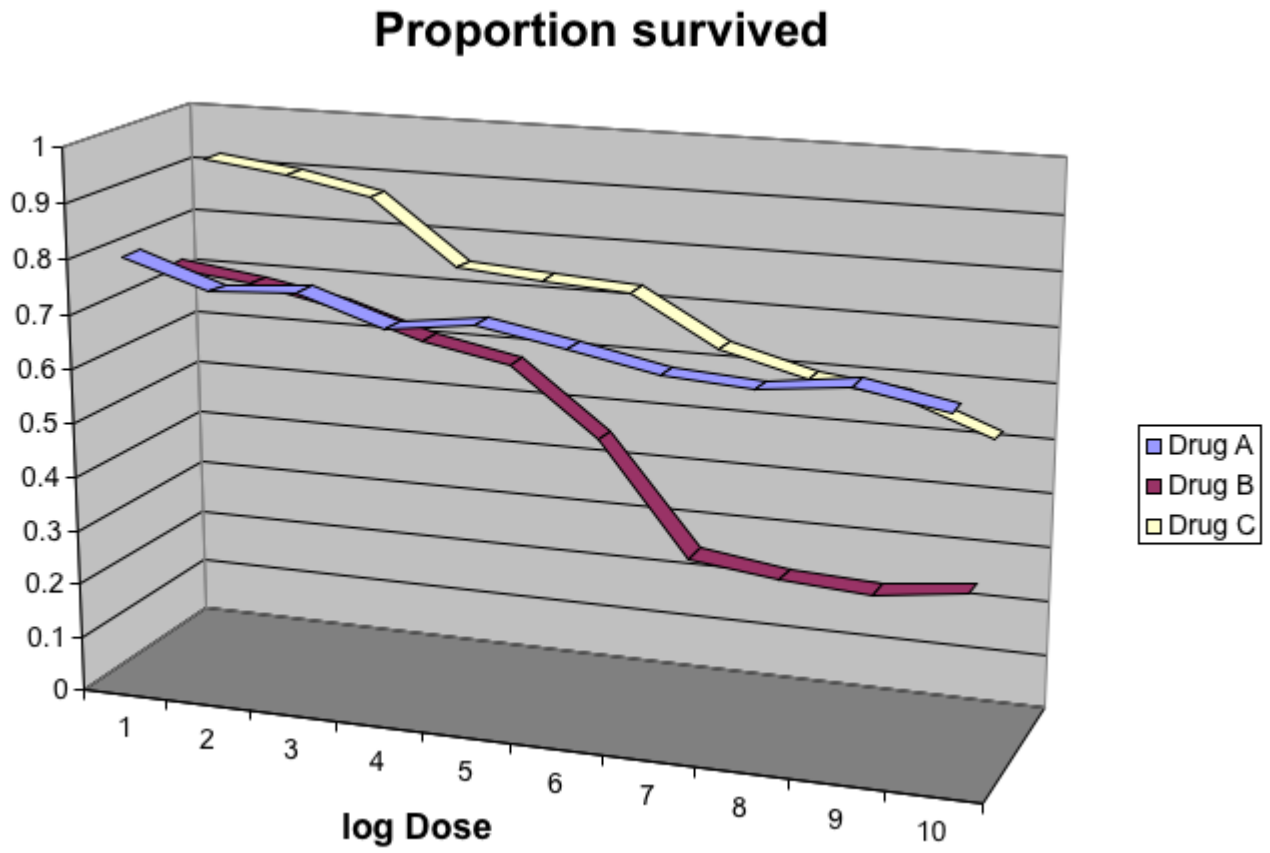
Adaptado de:
HEALY, K. **Data Visualization: A Practical Introduction.** Princeton University Press, 2019.

Adaptado de:
HEALY, K. **Data Visualization: A Practical Introduction.** Princeton University Press, 2019.

“Mau-gosto” ou problemas estéticos

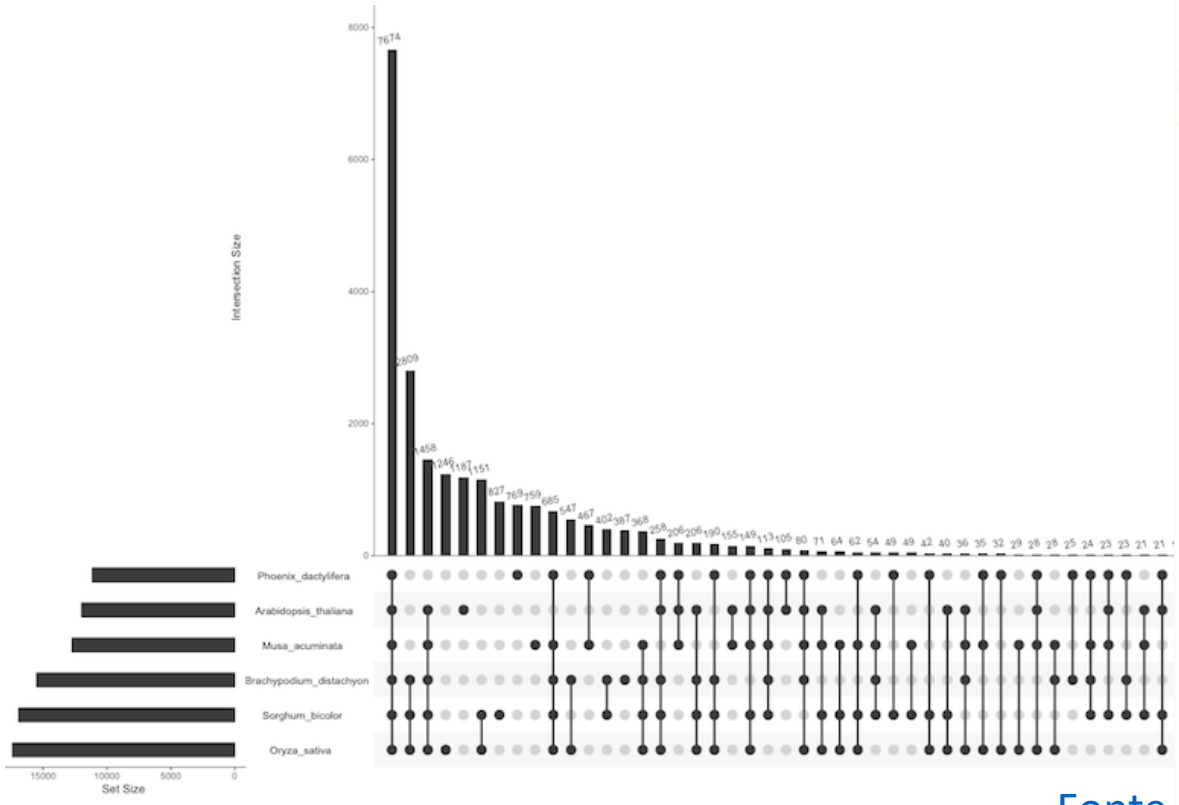


Fonte



Fonte

“Mau-gosto” ou problemas estéticos

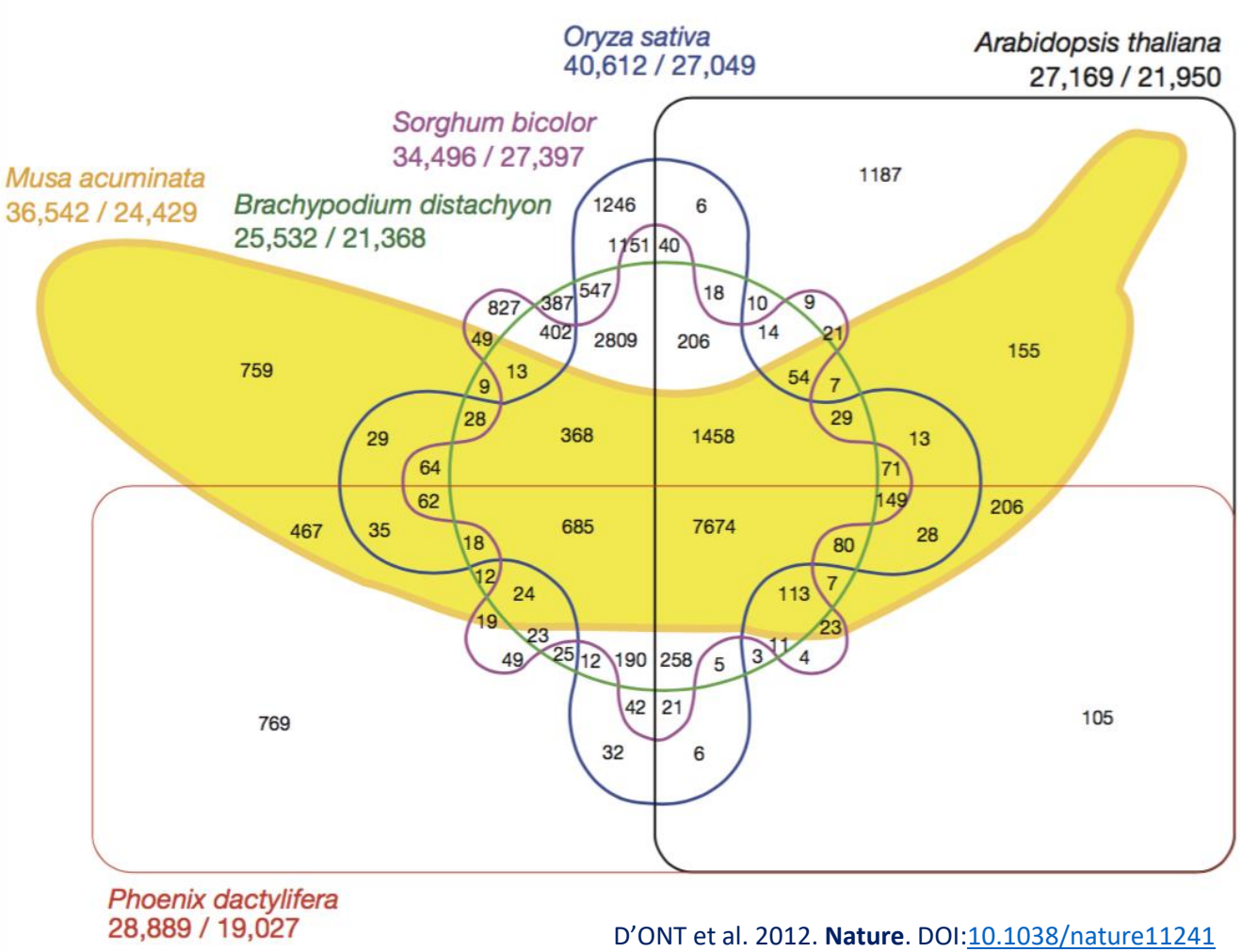


Fonte

LETTER

doi:10.1038/nature11241

The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants



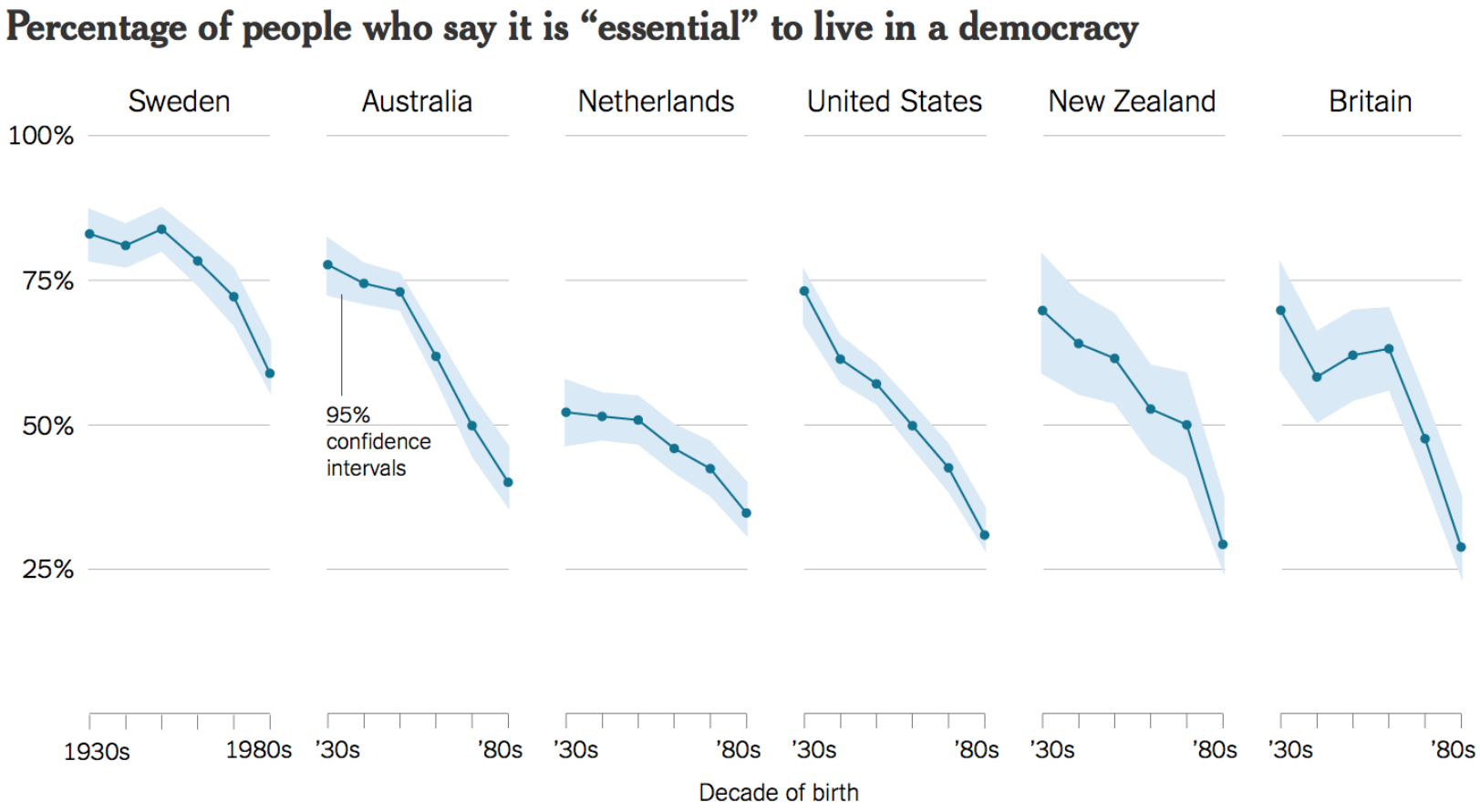
D’ONT et al. 2012. **Nature**. DOI:10.1038/nature11241

Problemas com os dados

- *"How Stable are Democracies? Warning Signs Are Flashing Red"*

- Reportagem do New York Times publicada com base em um artigo da área de ciência política

- Eixo x: pessoas de diferentes idades respondendo à pergunta, e não as mesmas pessoas respondendo à pergunta em diferentes momentos

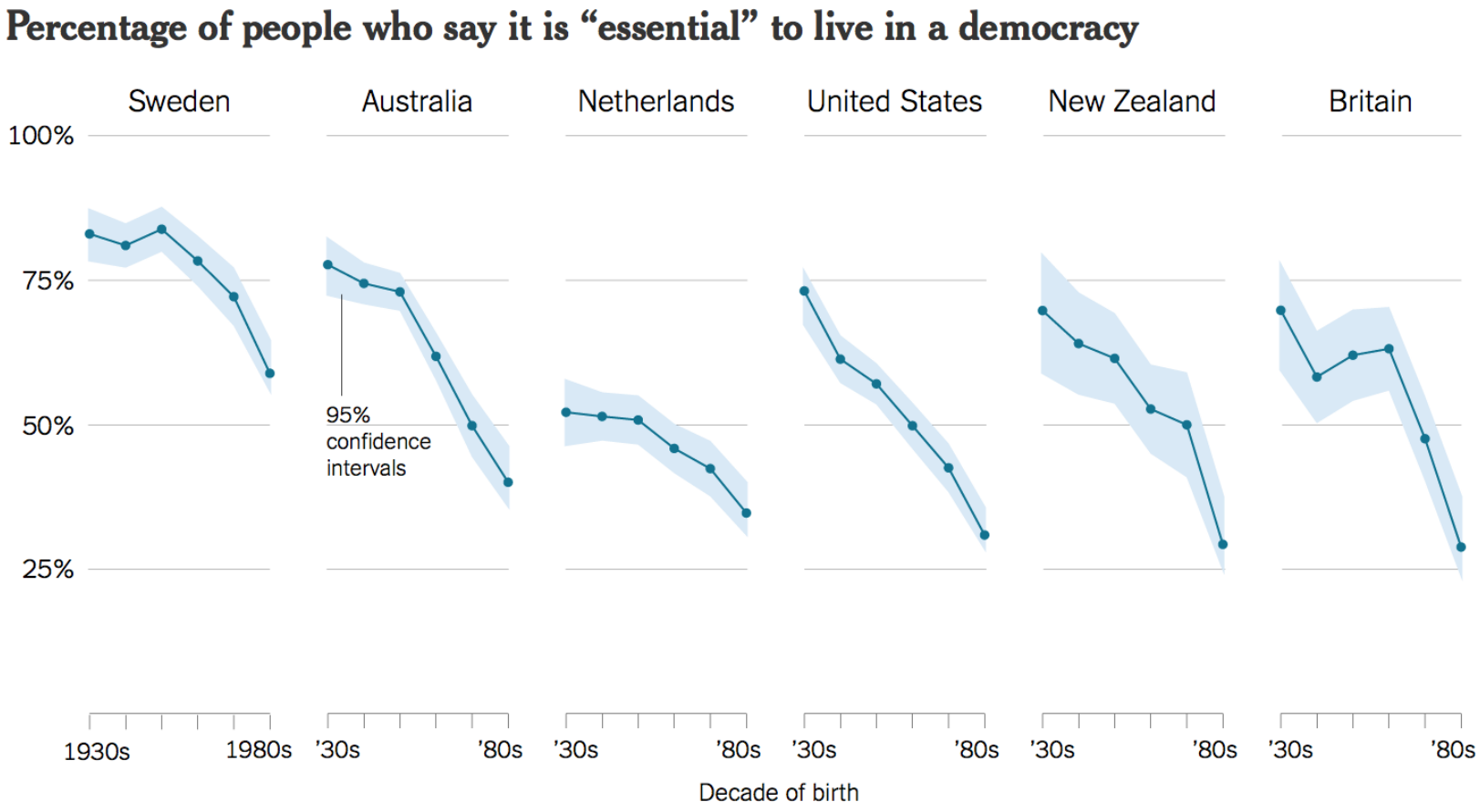


Problemas com os dados

- *"How Stable are Democracies? Warning Signs Are Flashing Red"*

- Gráfico sugere resposta de "Sim" ou "Não", mas as respostas reais estavam em escala:
  - 1: Não importante de forma alguma
  - 10: Absolutamente importante

- Gráfico representa somente as pessoas que responderam "10" (mesmo que "9" também seja considerar importante viver em uma democracia)



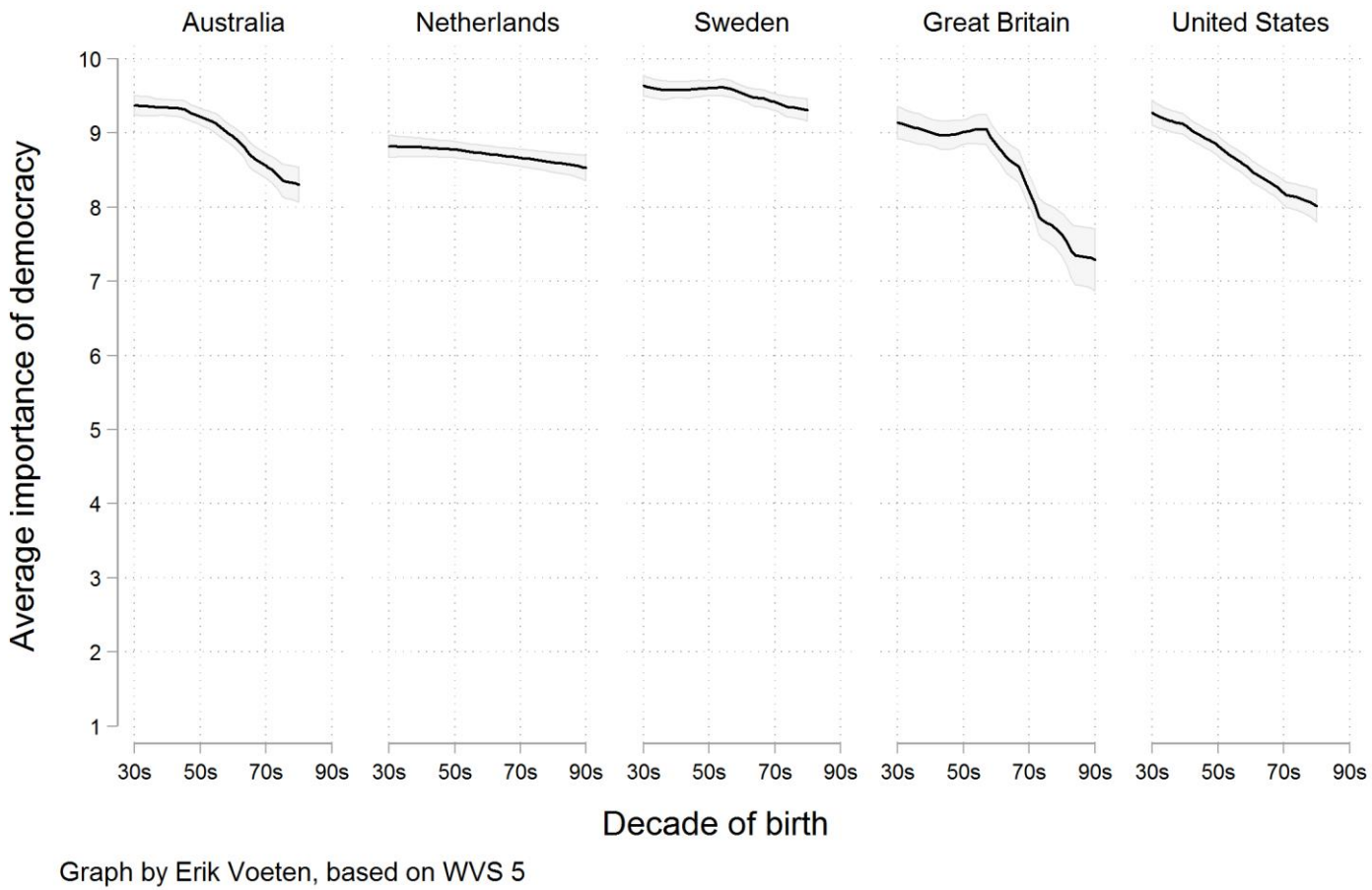
Source: Yascha Mounk and Roberto Stefan Foa, "The Signs of Democratic Deconsolidation," Journal of Democracy | By The New York Times

Problemas com os dados

- *"How Stable are Democracies? Warning Signs Are Flashing Red"*

- Ao utilizar a média das respostas, a situação aparenta ser menos alarmante

- Agora, observa-se a tendência média, e não somente a tendência do valor máximo



Adaptado de:
HEALY, K. **Data Visualization: A Practical Introduction.** Princeton University Press, 2019.

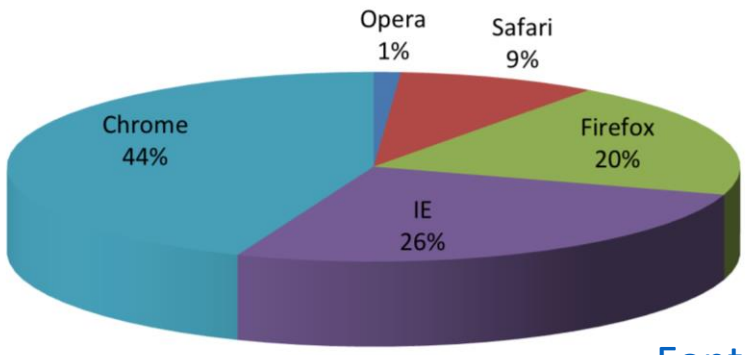
Problemas de percepção

- Território “nebuloso” entre dados e elementos estéticos
- Visualização de dados normalmente codifica números em linhas, formas e cores
- Nossa interpretação está condicionada à nossa capacidade de decodificar linhas, formas e cores, perceber formas geométricas e relações entre elementos

Problemas de percepção

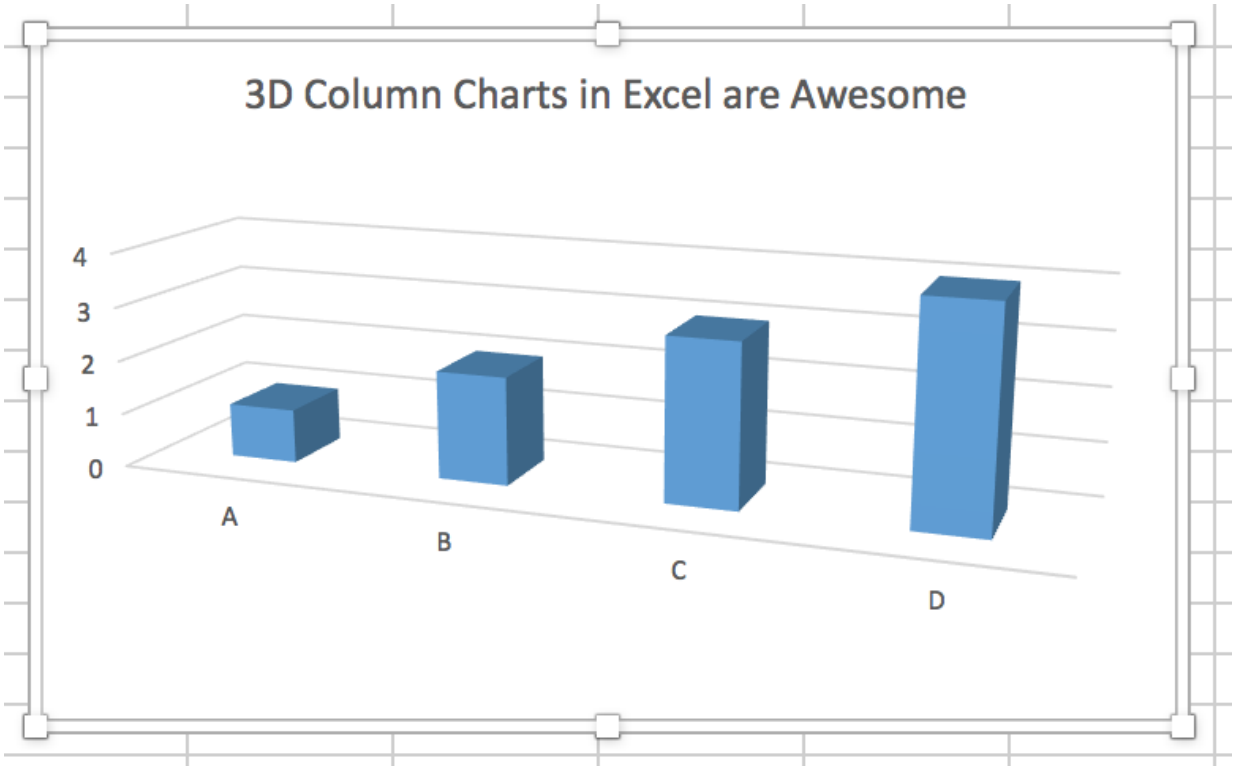
- A adição de dimensões desnecessárias torna as figuras confusas

- Principalmente quando a dimensão adicional não representa nenhuma informação ou variável



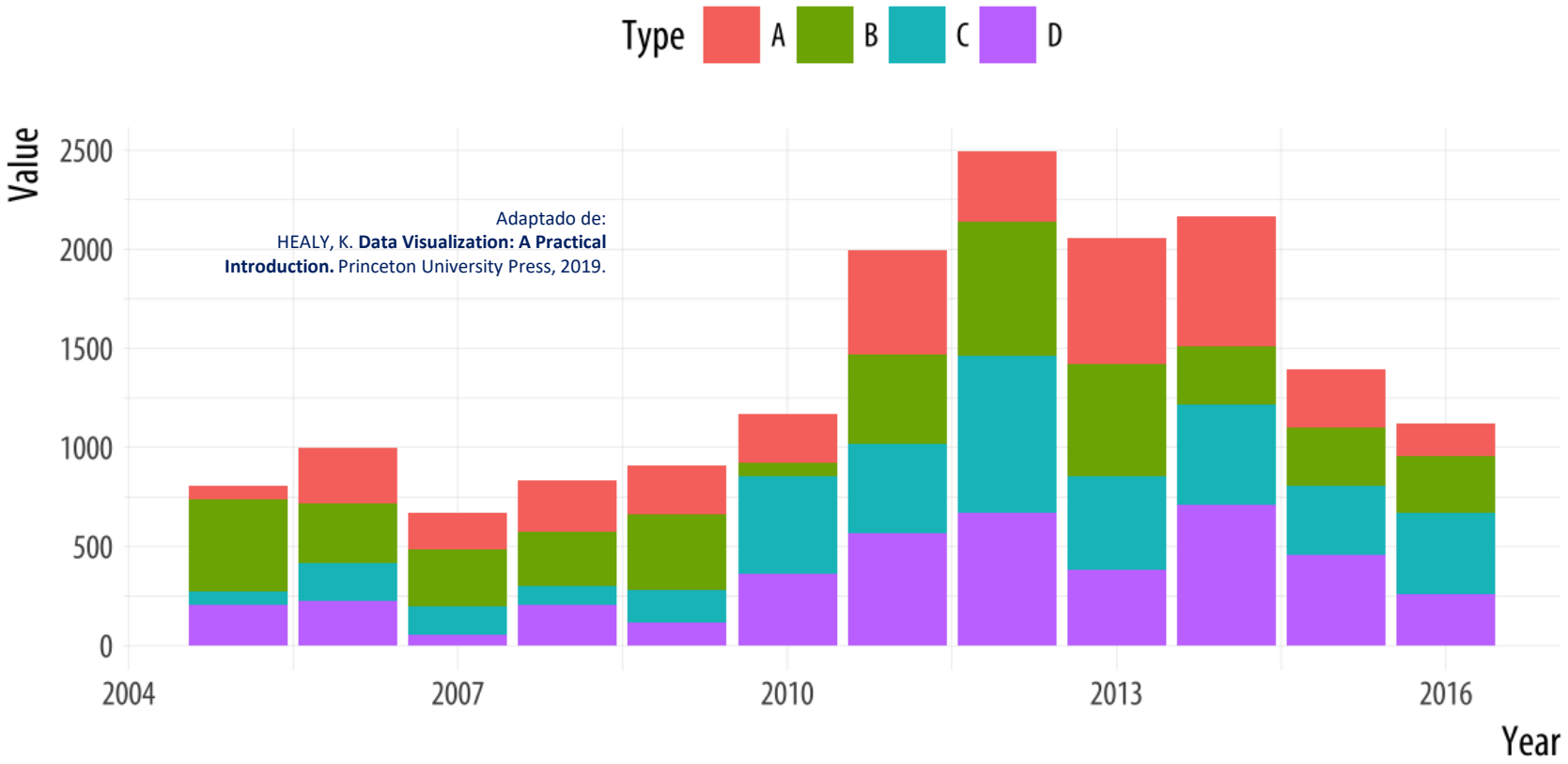
Fonte

A	1
B	2
C	3
D	4



Adaptado de:
HEALY, K. **Data Visualization: A Practical Introduction.** Princeton University Press, 2019.

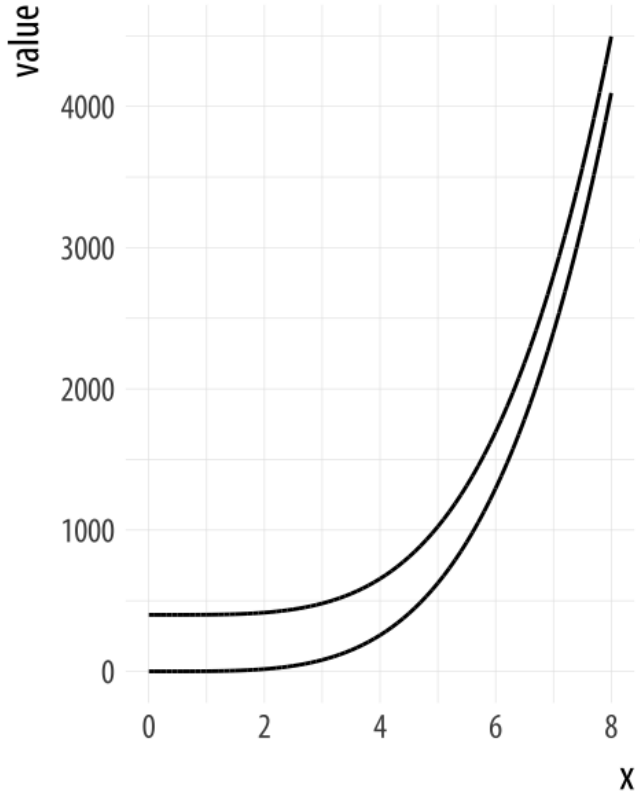
Problemas de percepção



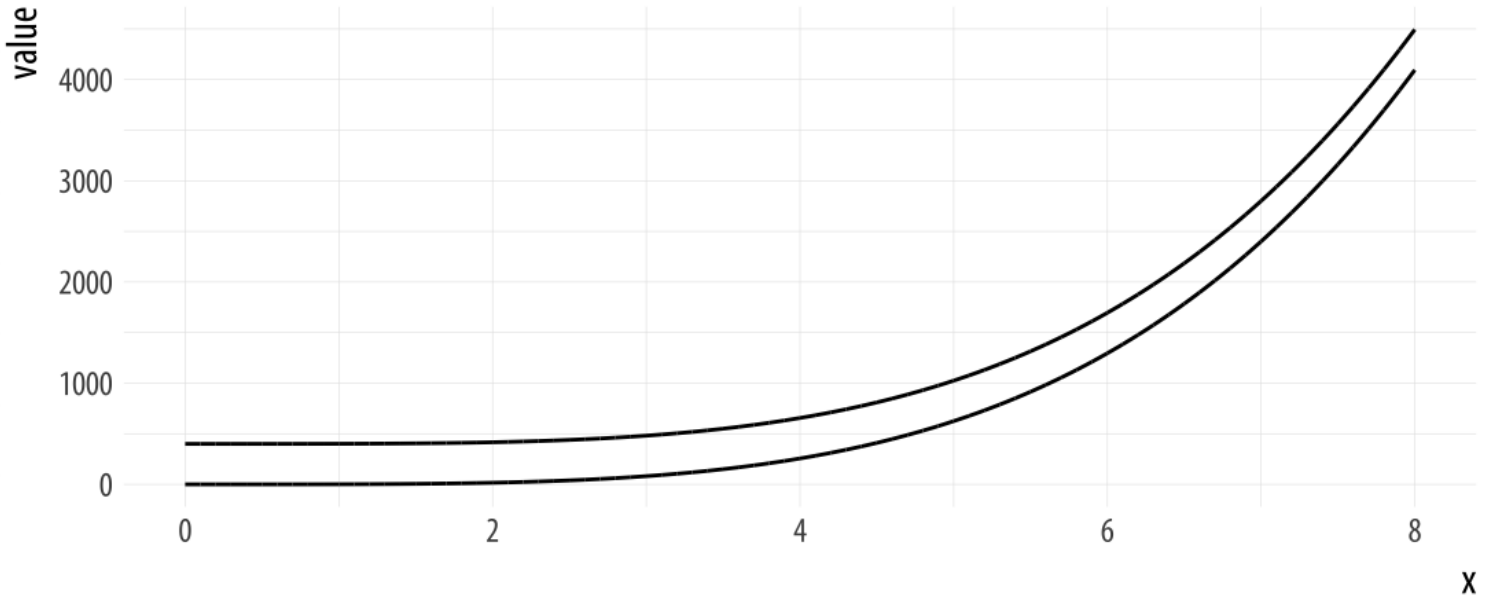
- Comparações relativas necessitam de uma linha de base estável

- Muito mais fácil comparar a tendência geral ou a tendência D do que as tendências A, B ou C

Problemas de percepção



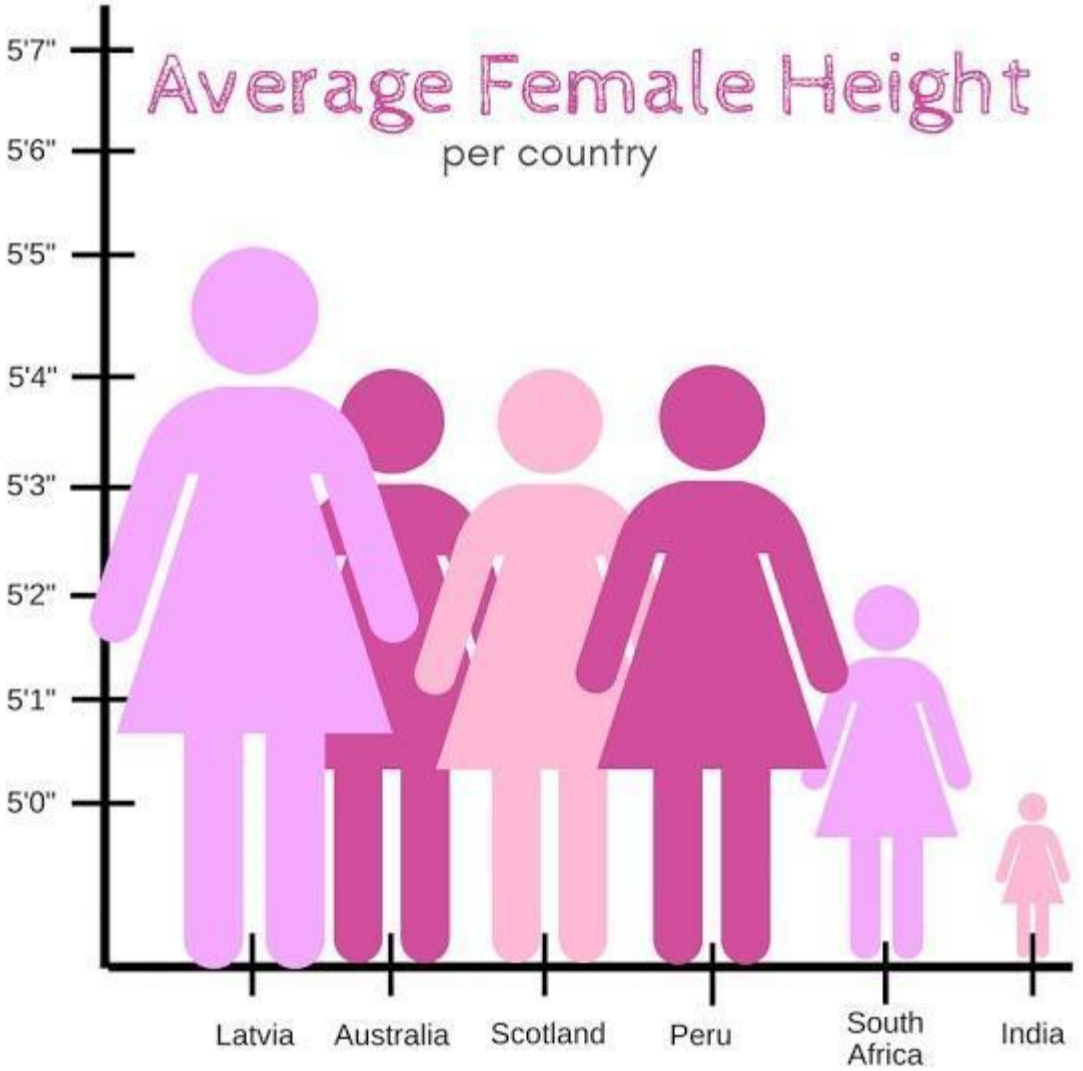
Adaptado de:
HEALY, K. **Data Visualization: A Practical Introduction.** Princeton University Press, 2019.



- A aparente convergência observada na primeira figura é resultado somente da proporção entre altura e largura da imagem
- Ambas as linhas são equidistantes

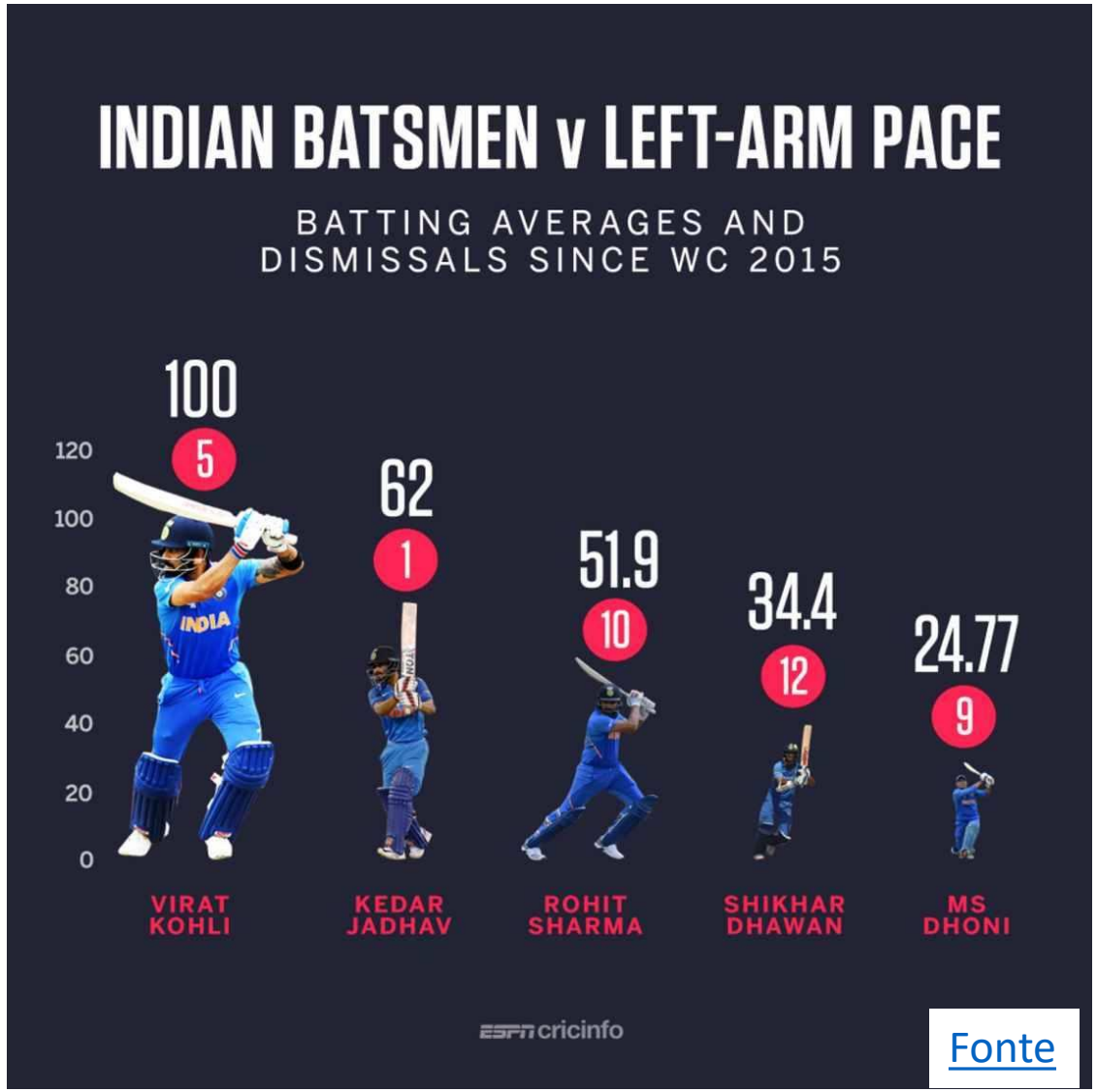
“Mau-gosto”, problemas de dados e percepção normalmente aparecem de forma combinada...

Fonte

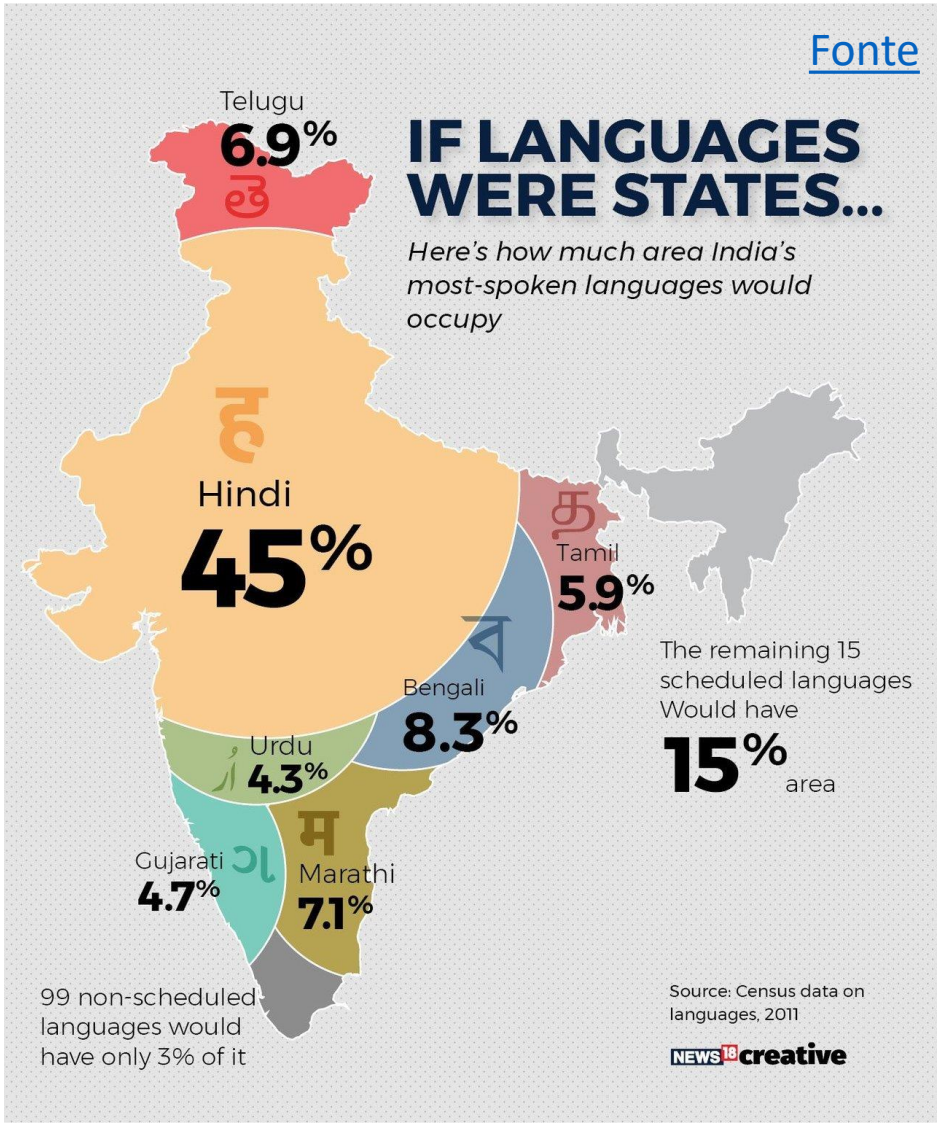


Fonte

“Mau-gosto”, problemas de dados e percepção normalmente aparecem de forma combinada…



Fonte



Percepção visual

- Nosso sistema visual faz com que algumas coisas sejam mais facilmente observadas que outras

- Ilusões de ótica demonstram que a percepção visual não é simplesmente resultado de uma relação direta entre entrada de dados visuais e produção de uma representação destes dados no cérebro

- Para que algumas tarefas visuais sejam bem desempenhadas, outras tarefas se tornam mais limitadas

Limites, contrastes e cores

Adaptado de:
HEALY, K. **Data Visualization: A Practical Introduction.** Princeton University Press, 2019.

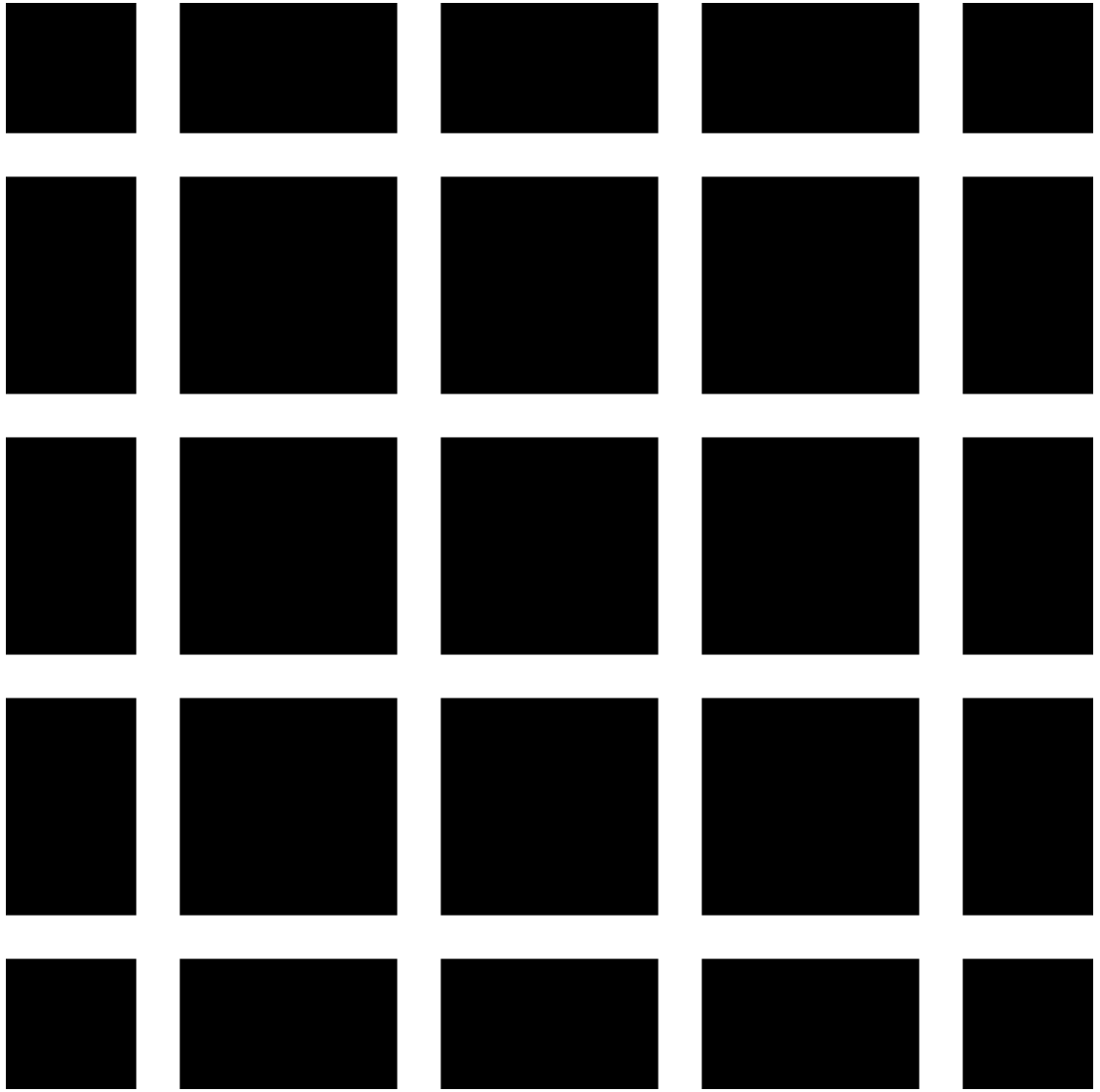
- Bandas de Mach: nosso sistema visual se baseia mais nas diferenças relativas de luminosidade entre as barras ao invés da luminosidade absoluta

- Quando as barras se tocam, as áreas escuras aparentam ser mais escuras e as áreas claras aparentam ser mais claras

Limites, contrastes e cores

- Grelha de Hermann: manchas fantasmas surgem nas intersecções das linhas

- As manchas desaparecem quando se observa diretamente a intersecção



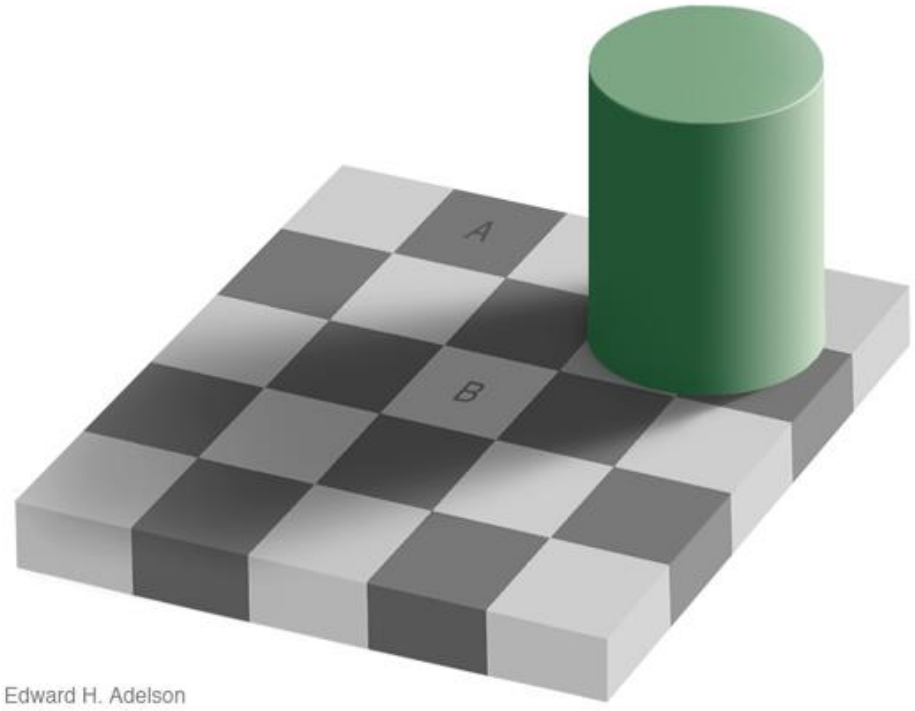
Fonte

Limites, contrastes e cores

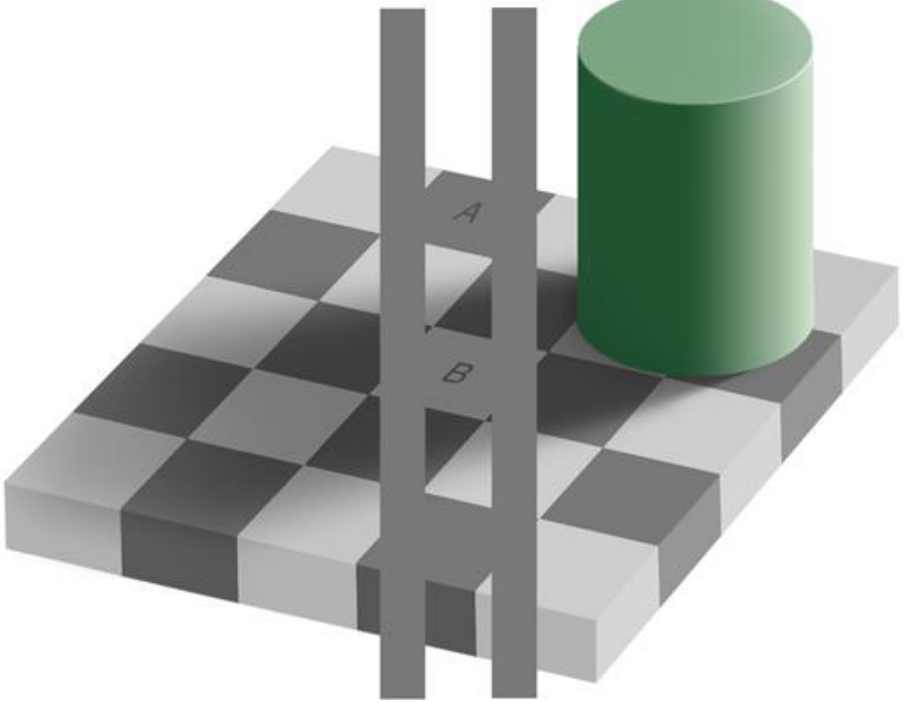
- Nosso sistema visual é atraído para limites e intersecções: contraste e luminosidade são determinados em termos relativos e não absolutos

- Um mesmo tom de cinza e diferenças de luminosidade são percebidos de forma diferente dependendo do contexto:
  - Mais fácil distinguir entre tons mais escuros
  - Mais fácil distinguir entre tons claros em fundo claro
  - Mais fácil distinguir entre tons intermediários em fundo escuro
  - Diferenças de percepção em função da sombra gerada por objetos adjacentes

Limites, contrastes e cores



Fonte



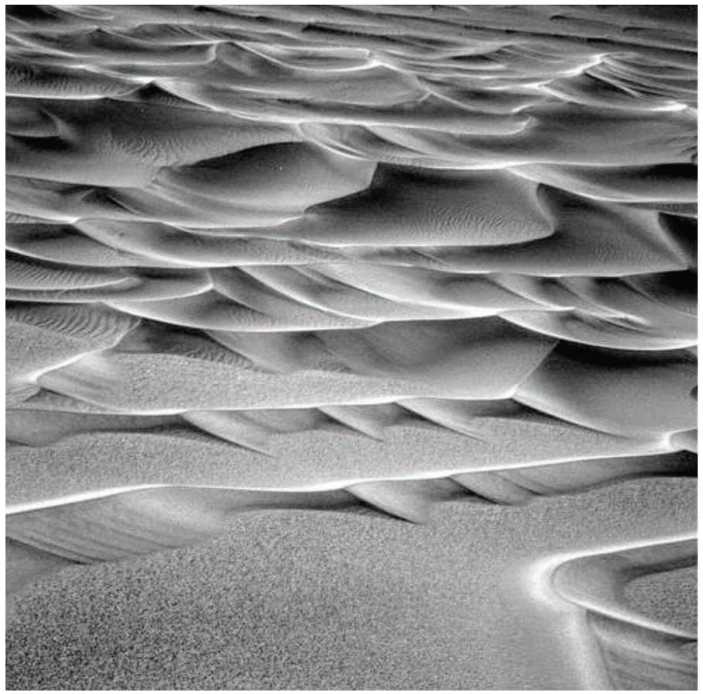
Fonte

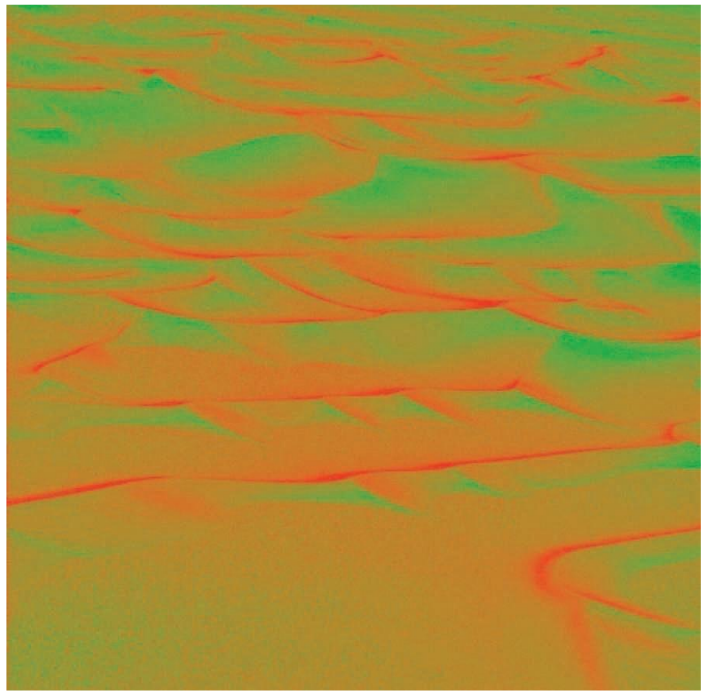
- Tabuleiro de Xadrez de Adelson: diferença de percepção do tom de cinza dos quadrados em função do contexto

- Sistema visual desenvolvido para perceber objetos em seu contexto e ambiente, e não para atribuir tonalidades de cinza

Limites, contrastes e cores

- Maior facilidade de perceber contraste em intersecções e regiões adjacentes em imagens monocromáticas ao invés de imagens coloridas





Adaptado de:
HEALY, K. **Data Visualization: A Practical Introduction.** Princeton University Press, 2019.

Limites, contrastes e cores

- Uso de cores em visualização de dados traz dificuldades adicionais, principalmente pela percepção relativa de luminosidade

- Componentes adicionais da cor (HCL):
  - Matiz: comprimento de onda dominante da luz refletida pelo objeto observado, geralmente o que nos referimos ao falar de cores (“vermelho”, “azul”, “verde”)

  - Saturação: intensidade ou pureza da cor. Quanto maior a saturação, menor a presença de cinza na composição da cor

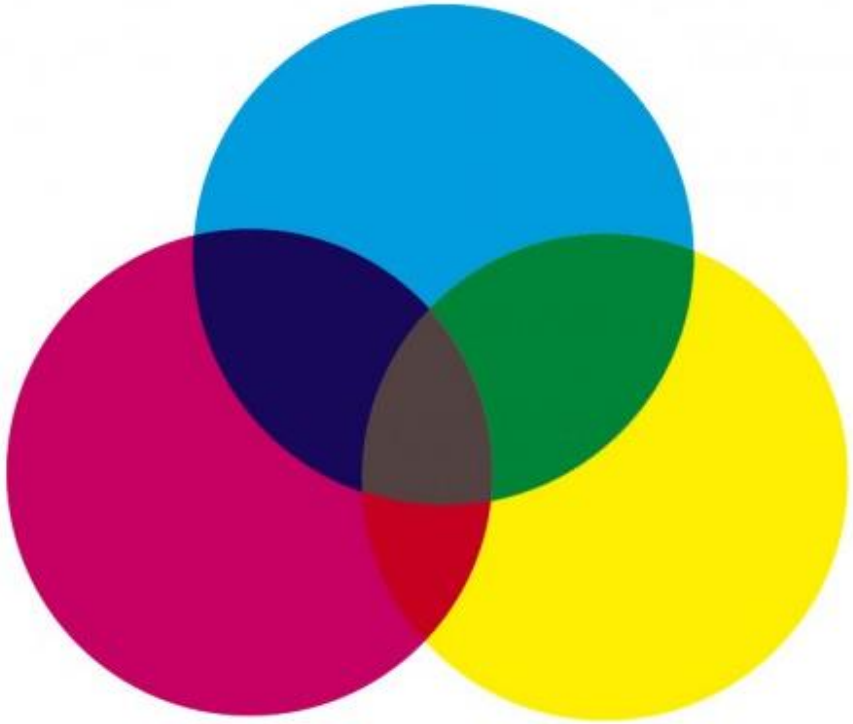
Limites, contrastes e cores

- Componentes da cor (HCL):

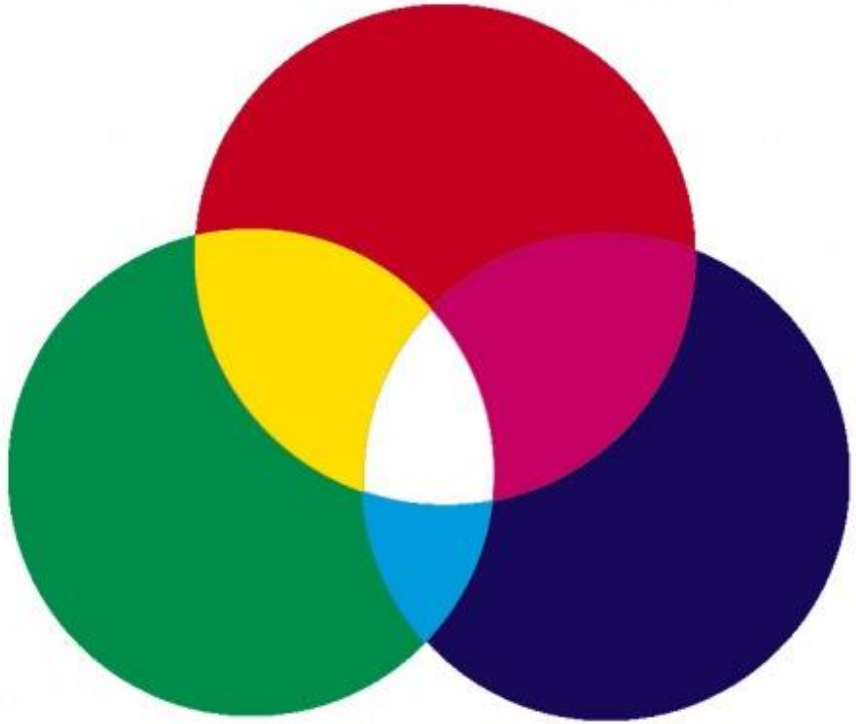
Fonte

Limites, contrastes e cores

- Diferentes sistemas e modelos de cor para gerar cores observadas em telas e superfícies de impressão



Subtractive color (CMYK)



Additive Color (RGB)

Fonte

Limites, contrastes e cores

- Ao utilizar cores numa figura, estamos mapeando variáveis numéricas ou categóricas

- Variável numérica (0, 1, 2, 3, 4, 5), sendo 0 o menor valor. Valores equidistantes entre si (incremento de 1)

- Primeira tendência “intuitiva”: utilizar cores equidistantes (em termos numéricos) para cada um dos valores da variável

Limites, contrastes e cores

- Entretanto, nossa percepção visual não é uniforme dentre o universo de cores possíveis (por exemplo, a faixa de saturação percebida depende da luminosidade)

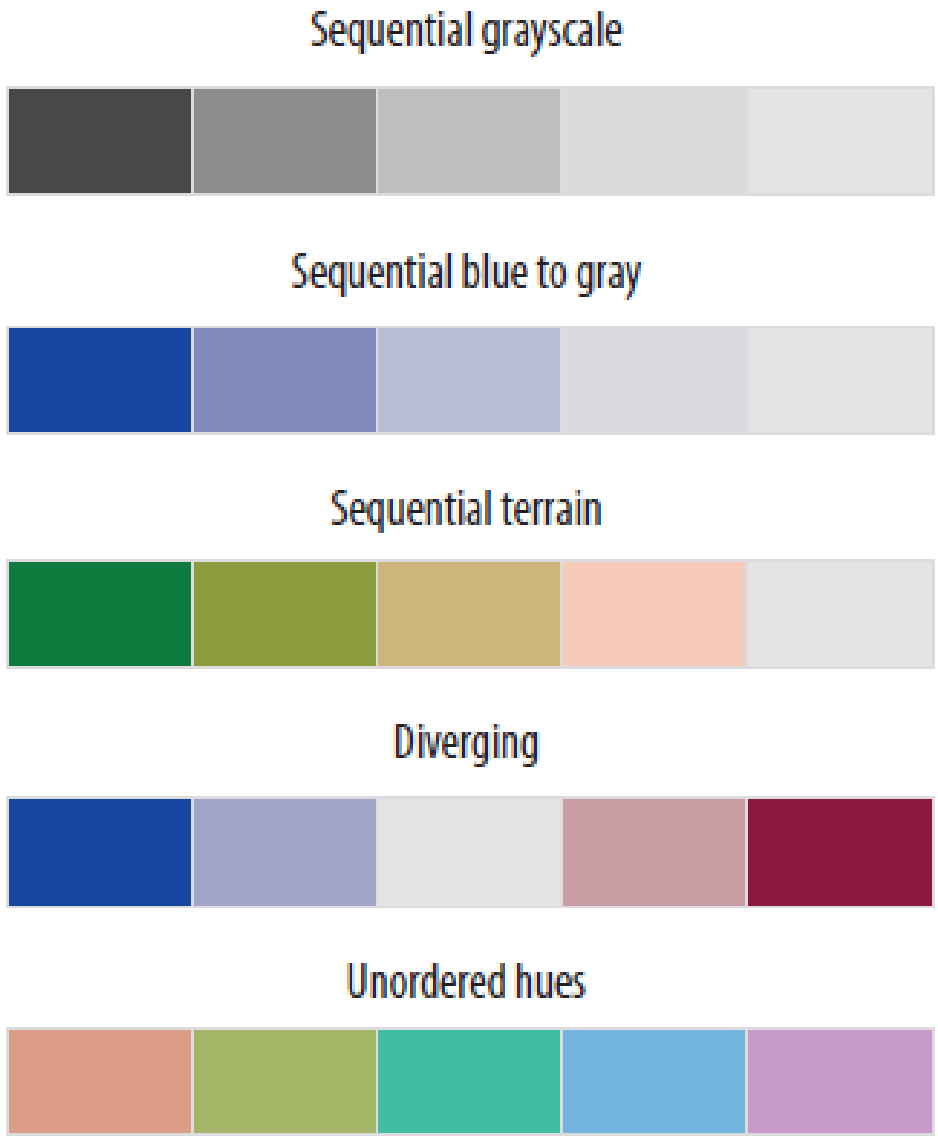
- As cores não seriam percebidas como equidistantes no gráfico

- Gradiente utilizando cores erradas: valores equidistantes poderiam ser percebidos de forma diferente pelo observador (por exemplo, diferença entre 0 e 1 poderia ser percebida de forma mais intensa que a diferença entre 3 e 4)

Limites, contrastes e cores

- Escolher cores priorizando uniformidade de percepção, e não uniformidade da correspondência entre variável e decodificação da cor dentro de um sistema

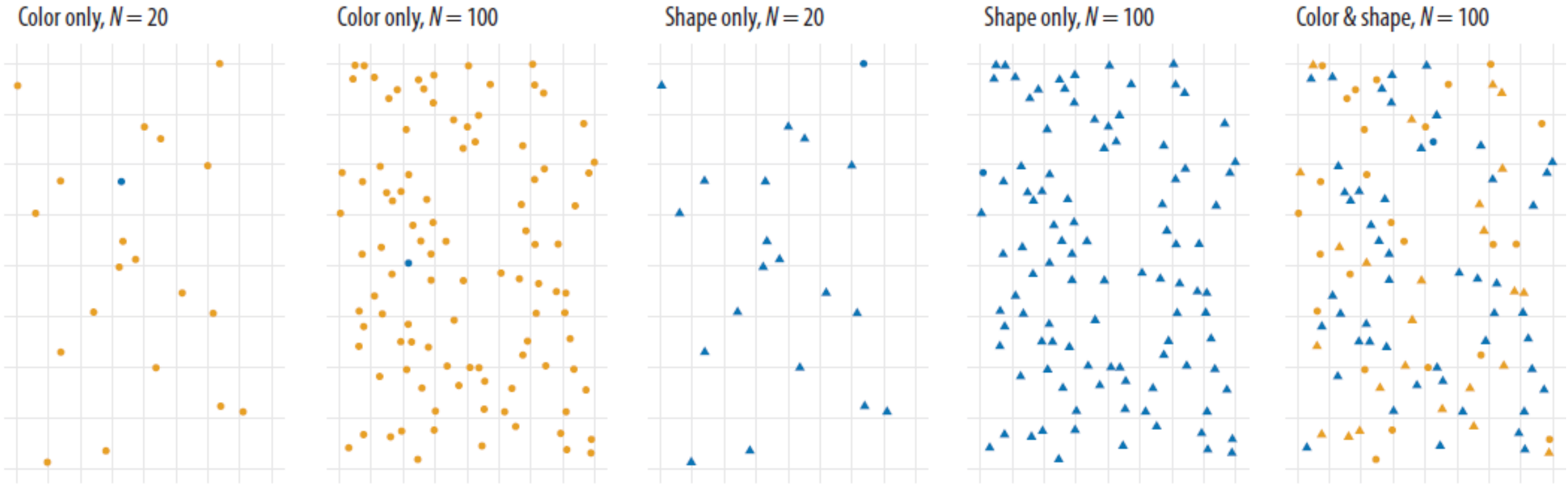
- Salvo exceções pontuais, não é desejável que uma cor domine a percepção em relação às outras



Adaptado de:
HEALY, K. **Data Visualization: A Practical Introduction.** Princeton University Press, 2019.

Processos pré-atentivos: o que “salta aos olhos”

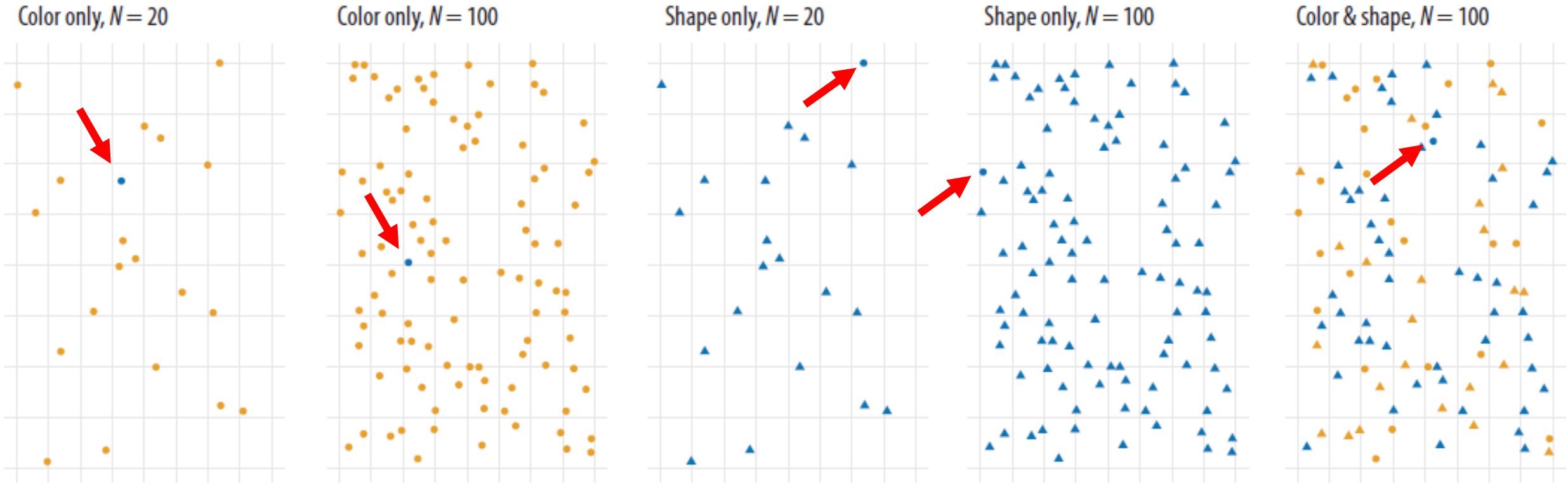
- Onde está o círculo azul?



Adaptado de:
HEALY, K. **Data Visualization: A Practical Introduction.** Princeton University Press, 2019.

Processos pré-atentivos: o que “salta aos olhos”

- Alguns tipos de canais de codificação de informação visual são mais facilmente percebidos que outros



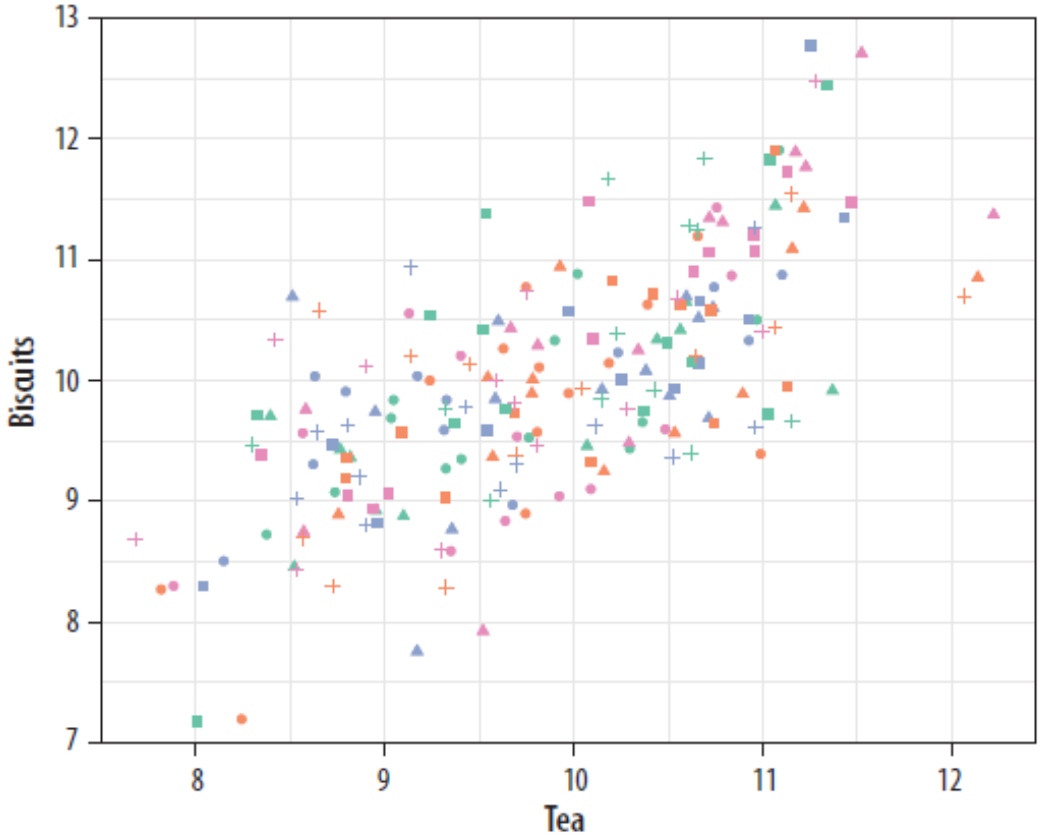
Adaptado de:
HEALY, K. **Data Visualization: A Practical Introduction.** Princeton University Press, 2019.

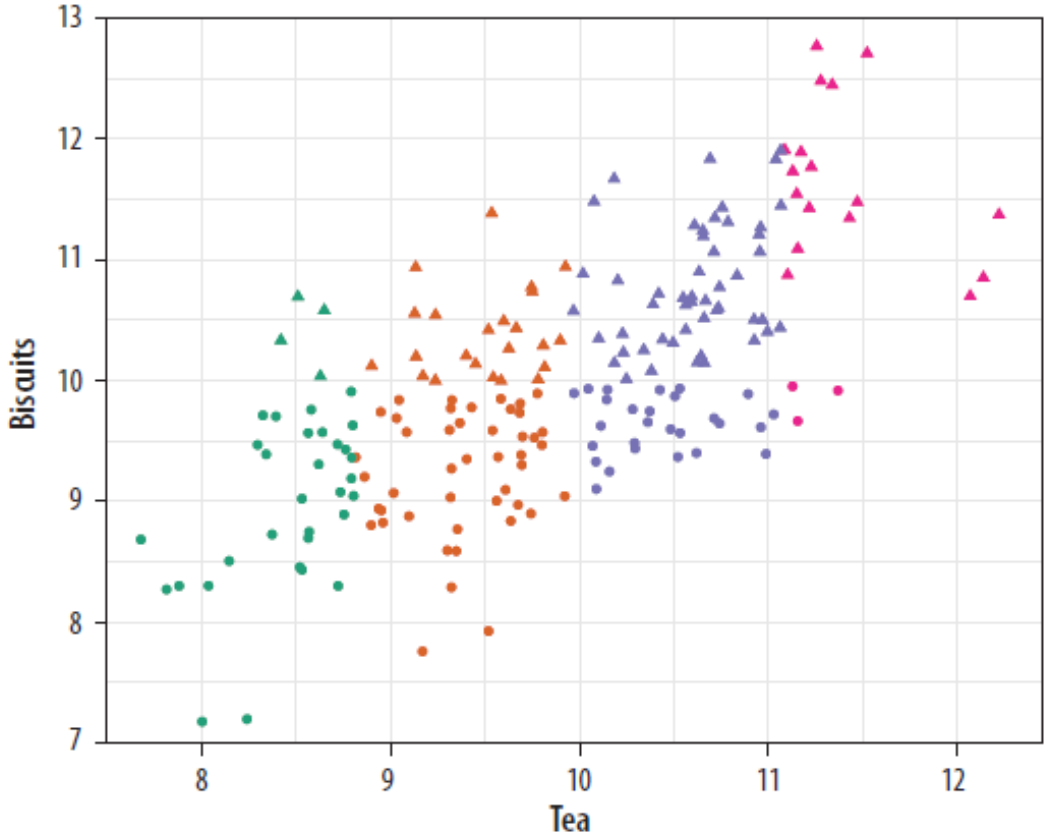
Processos pré-atentivos: o que “salta aos olhos”

- Processos pré-atentivos ocorrem antes mesmo de conscientemente procurarmos por algo
- Círculo azul mais facilmente encontrado quando há somente diferença de cor do que de forma
- Se cor e forma são canais visuais que podem codificar informação, a cor “salta mais aos olhos” do que a forma
- Outros canais possíveis: tamanho, ângulo, comprimento, comprimento

Processos pré-atentivos: o que “salta aos olhos”

- Quanto mais canais diferentes a serem analisados, pior a performance na busca por informação

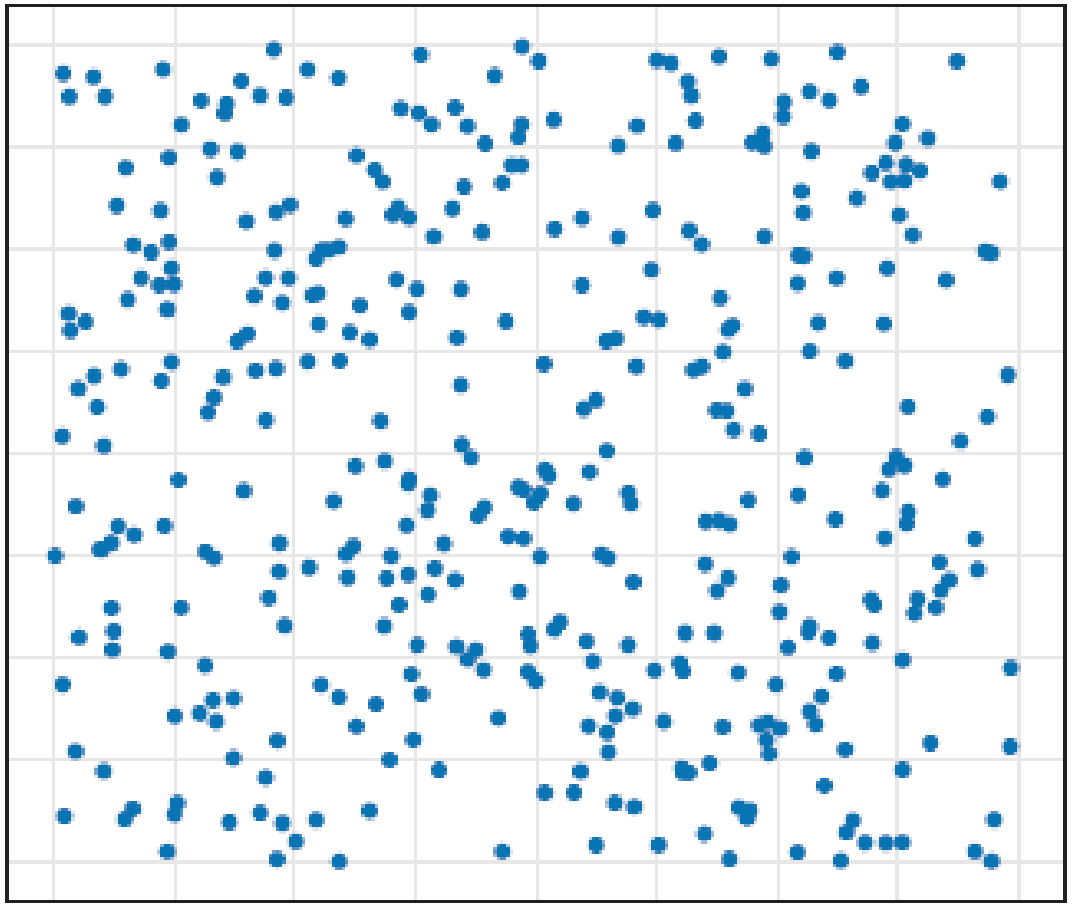




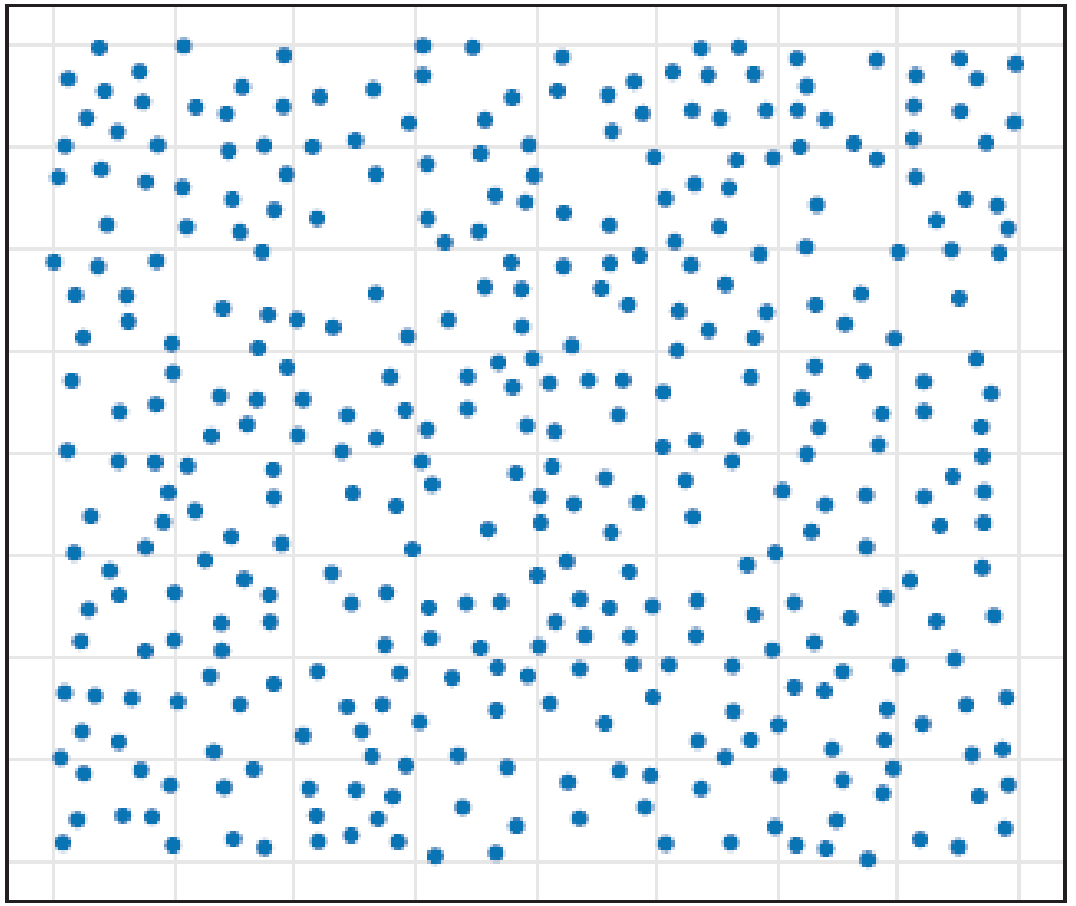
Adaptado de:
HEALY, K. **Data Visualization: A Practical Introduction.** Princeton University Press, 2019.

Qual padrão é mais aleatório?

Poisson



Matérn



Adaptado de:
HEALY, K. **Data Visualization: A Practical Introduction.** Princeton University Press, 2019.

Somos obcecados por encontrar estrutura nas coisas, até mesmo quando ela não existe...

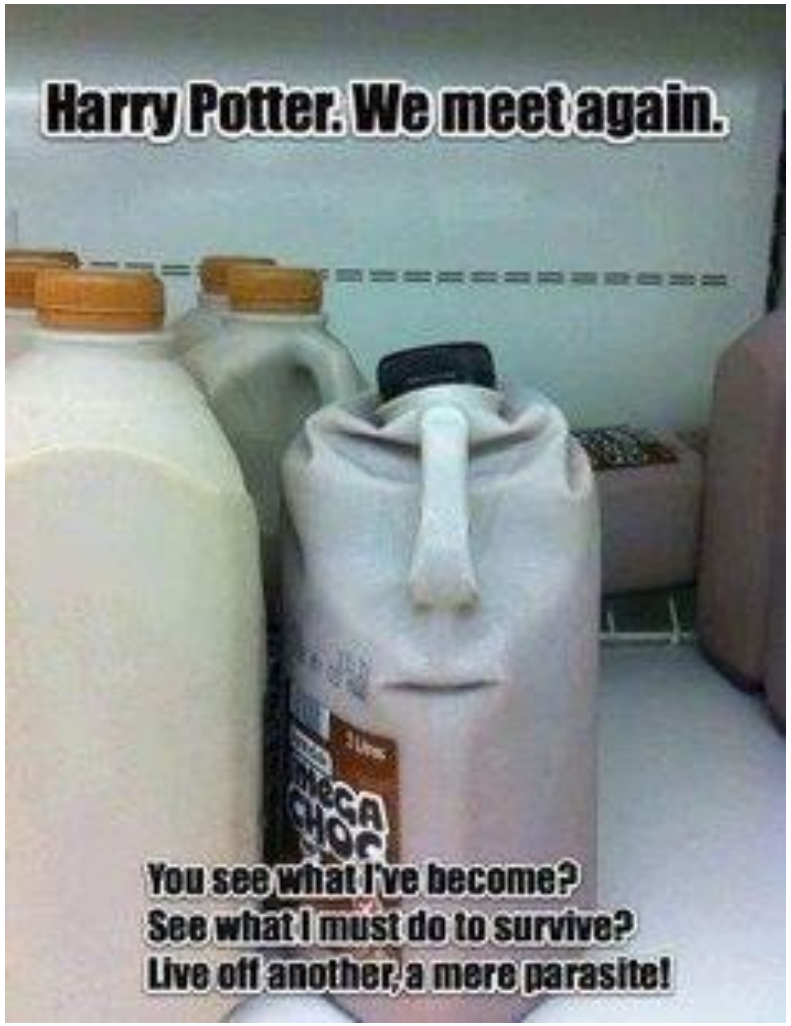
Fonte

Fonte

.

Somos obcecados por encontrar estrutura nas coisas, até mesmo quando ela não existe...

Fonte



Fonte

.

Princípios de Gestalt

- Exercício rápido: imagine uma mesa

- Provavelmente a figura que você imaginou é de uma mesa completa e não um tampo apoiado sobre quatro pés
- Formou-se a imagem completa, e não primeiro os pés, depois o tampo ou outros detalhes

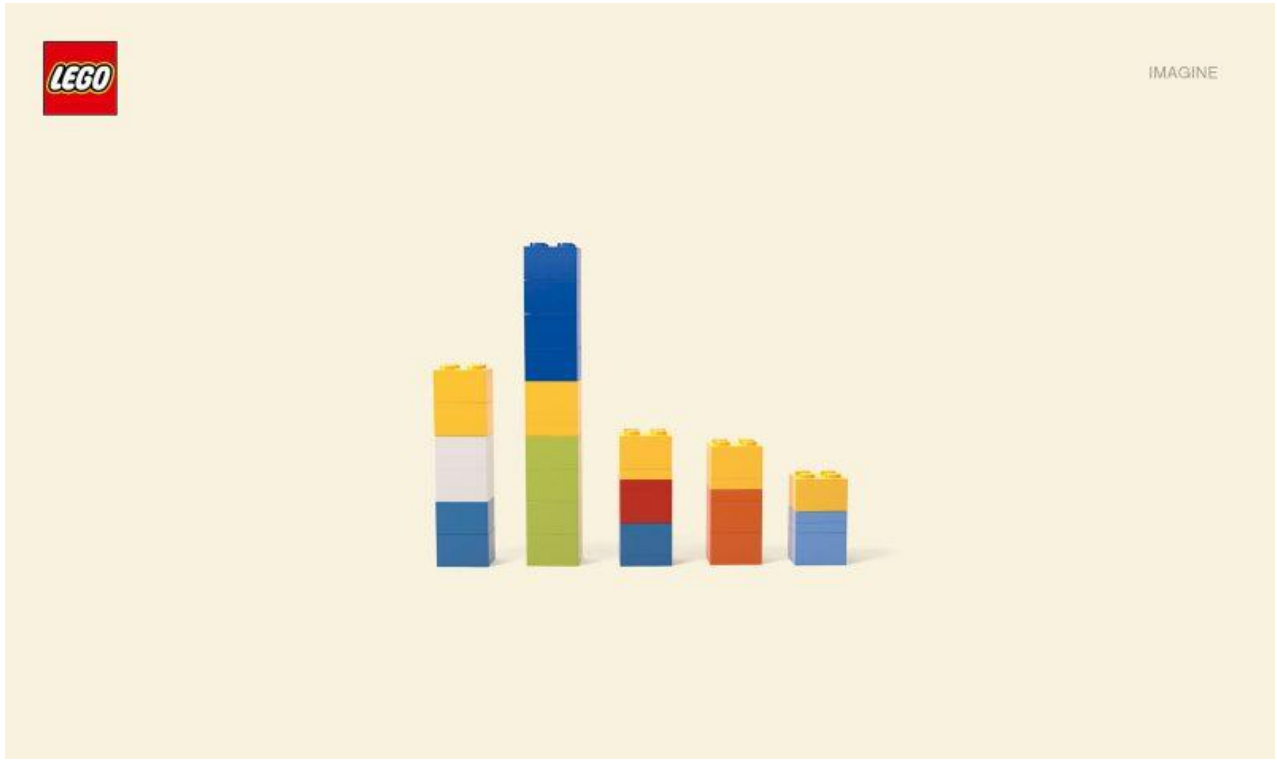
Gestalt (do alemão “figura” ou “forma”): campo de estudos dedicado à percepção das figuras pelo cérebro, defendendo que a percepção ocorre de maneira global e unificada, e não em partes isoladas

Princípios de Gestalt

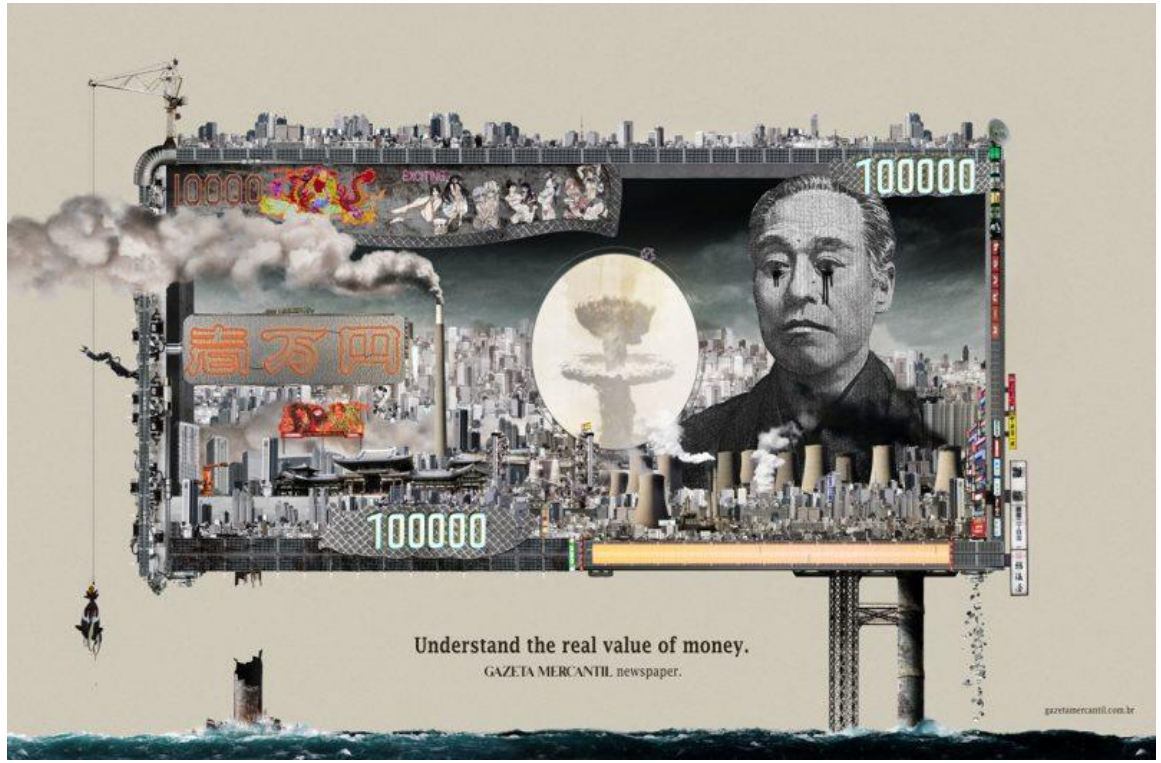
- Inferências sobre relacionamentos entre elementos visuais feitas a partir de informações visuais esparsas
- Tendência de inferir relações entre objetos além do que está estritamente visível
- Tendência de identificar agrupamentos, classificações e elementos que possam ser tratados como algo único, ou partes de algo único

Pregnância (Prägnanz)

- Simplicidade: quanto menos complexa for uma imagem, mais facilmente será assimilada



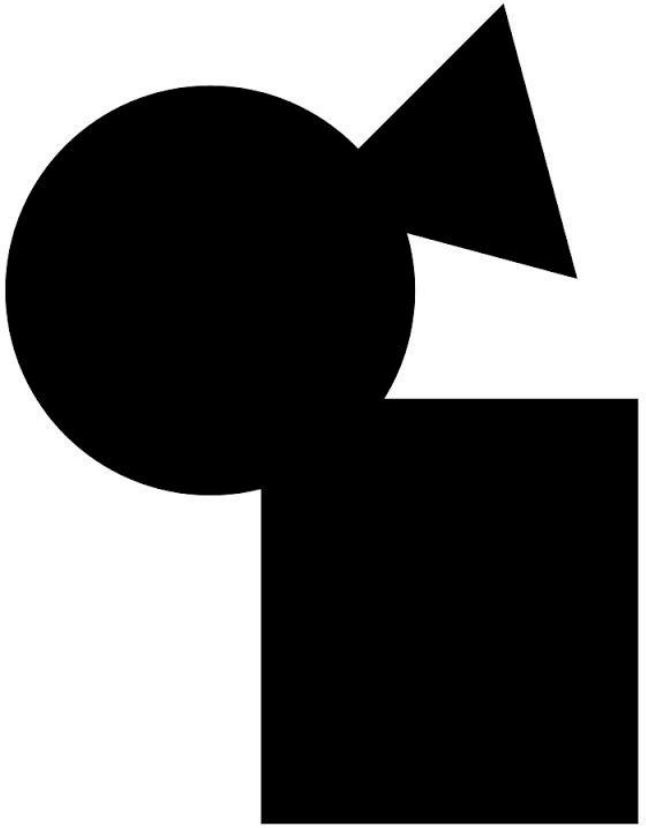
Fonte

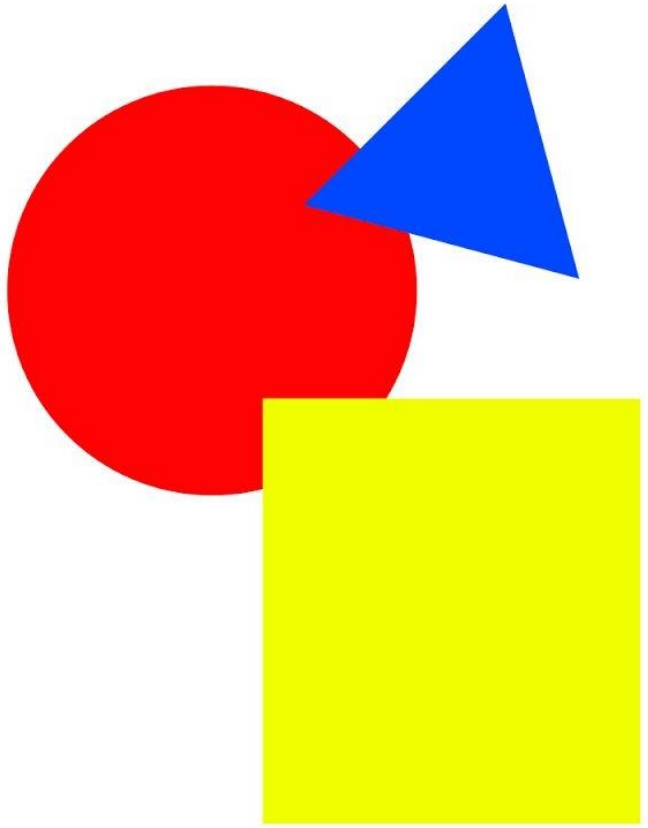


Fonte

Pregnância

- Interpretação de formas ambíguas da maneira mais simples possível





Fonte

Pregnância

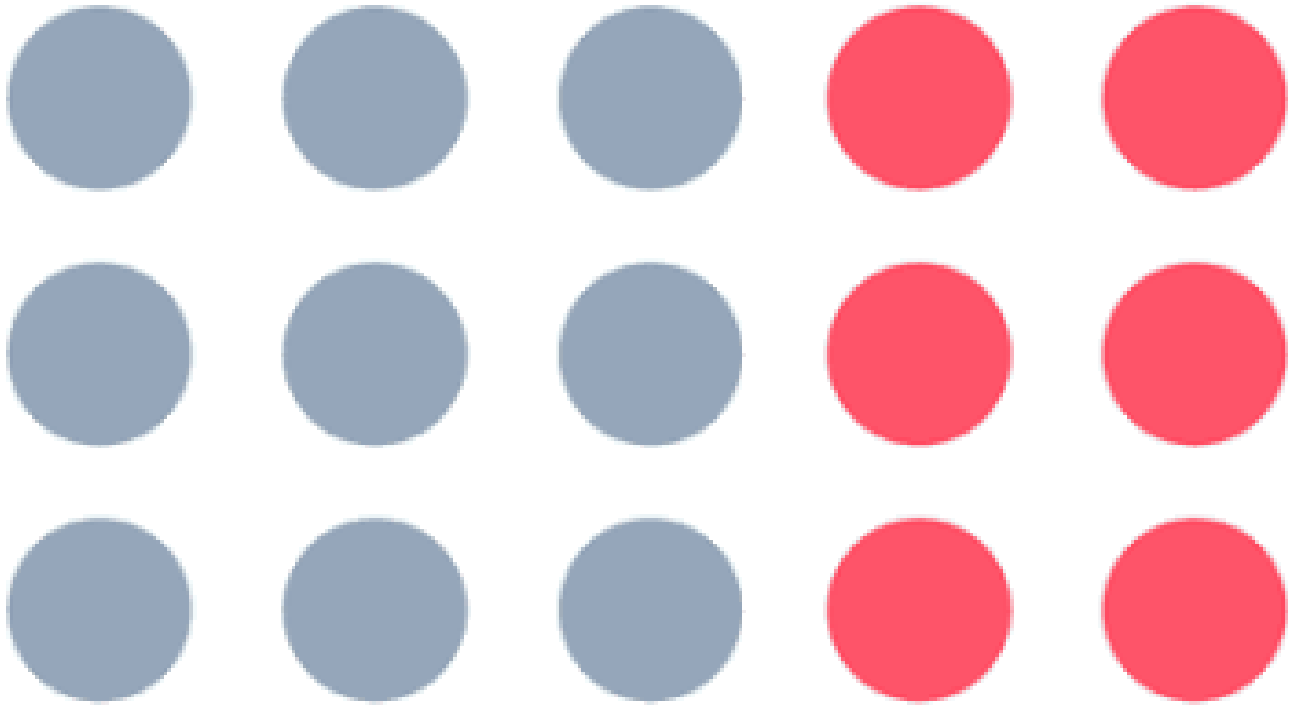
- Diferentes níveis de pregnância

Alta pregnância	A organização visual da forma do objeto interfere diretamente na rapidez e facilidade em compreendermos a leitura.
Média pregnância	*A organização visual da forma do objeto interfere diretamente na rapidez e facilidade em compreendermos a leitura.*
Baixa pregnância	

Fonte

Semelhança / Similaridade

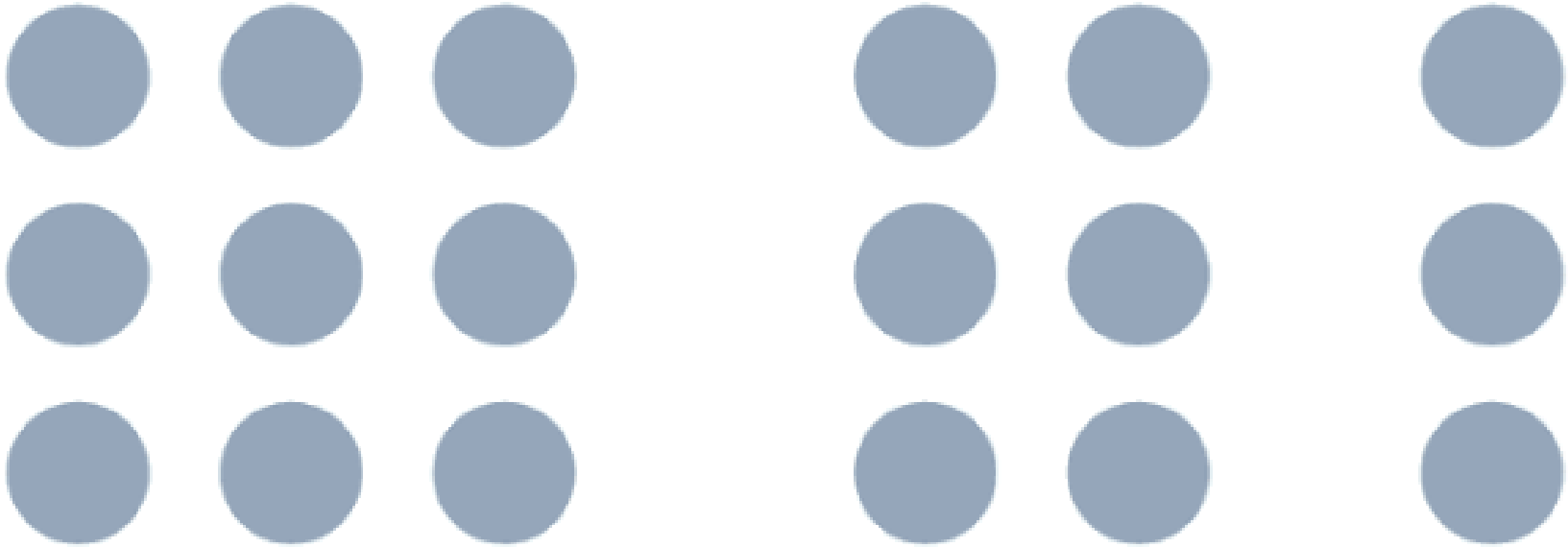
- Elementos similares (cor, textura, forma, dimensão) costumam ser vistos como agrupados ou relacionados



Fonte

Proximidade

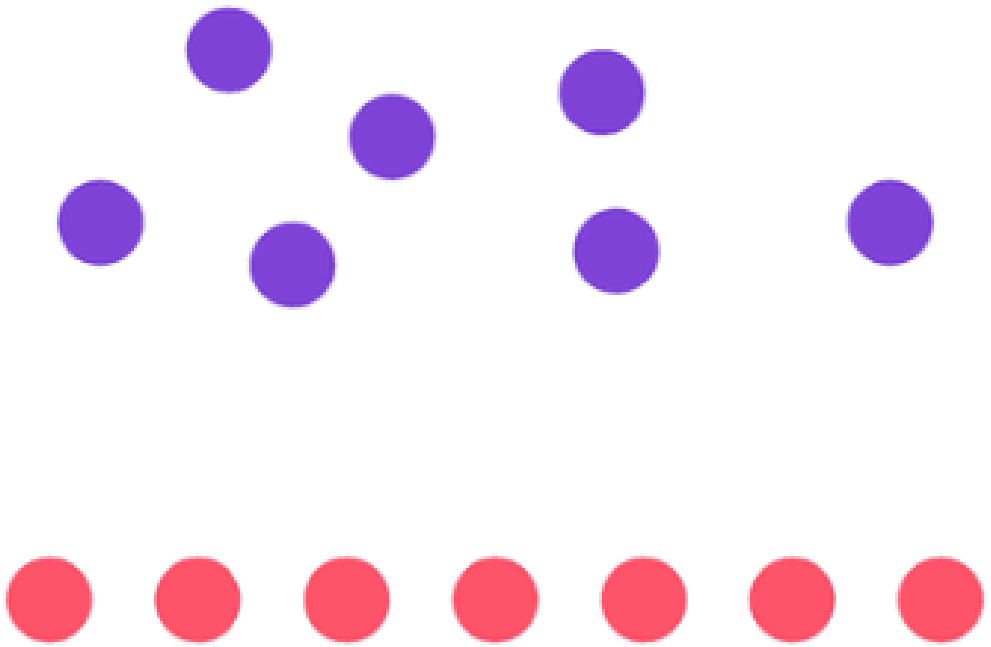
- Elementos próximos espacialmente costumam ser vistos como agrupados ou relacionados



Fonte

Continuidade

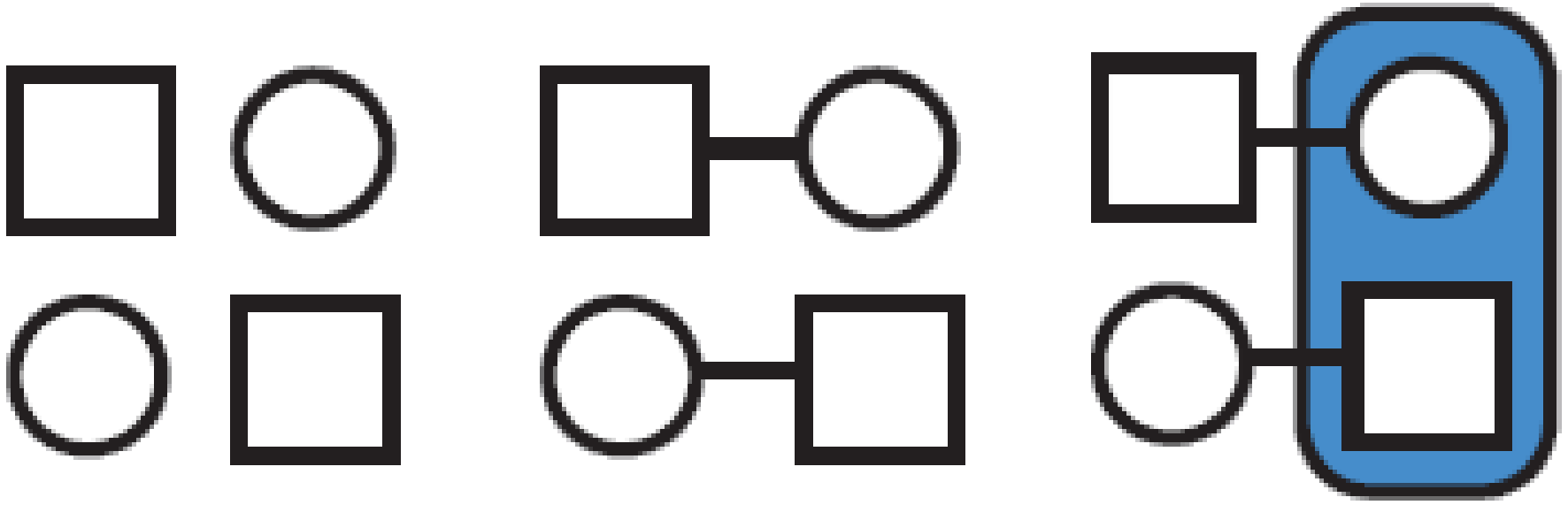
- Pontos que parecem interligados por linhas retas ou curvas são visualizados como uma unidade, e não linhas separadas



Fonte

Conexão

- Elementos visualmente conectados a outros aparentam estar relacionados



Adaptado de:
HEALY, K. **Data Visualization: A Practical Introduction.** Princeton University Press, 2019.

Fechamento

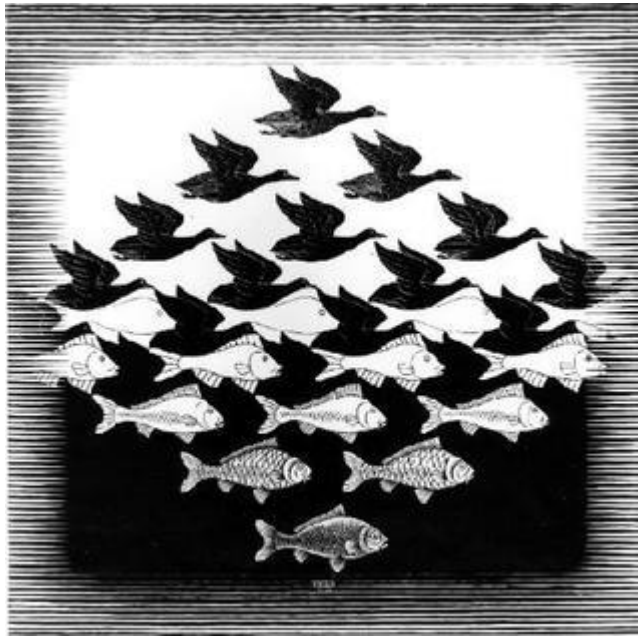
- Tendência de buscar o fechamento visual de imagens abertas, incompletas ou vazadas para atribuir significado

Fonte

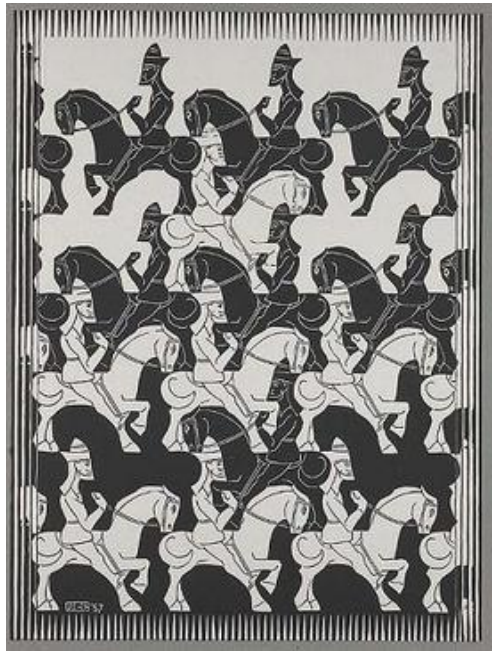
Fonte

Figura/fundo

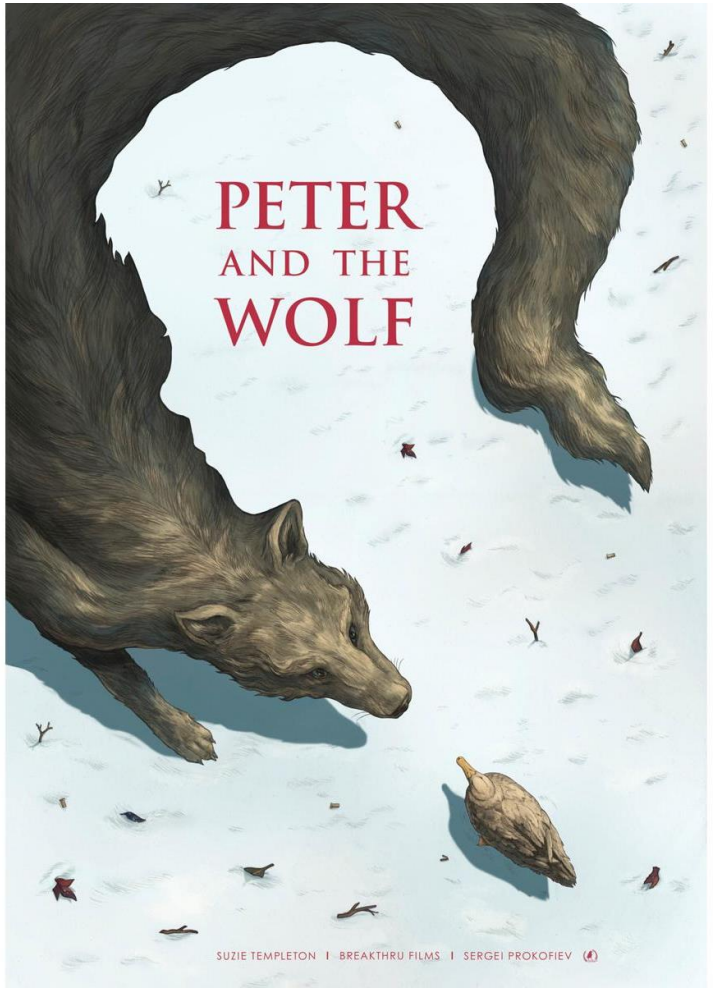
- Tendência do sistema visual de simplificar cenas com um objeto principal que olhamos (figura) e o restante (fundo)
- Em geral, áreas maiores são interpretadas como fundo e as menores como figura



Sky and Water I (M. C. Escher)



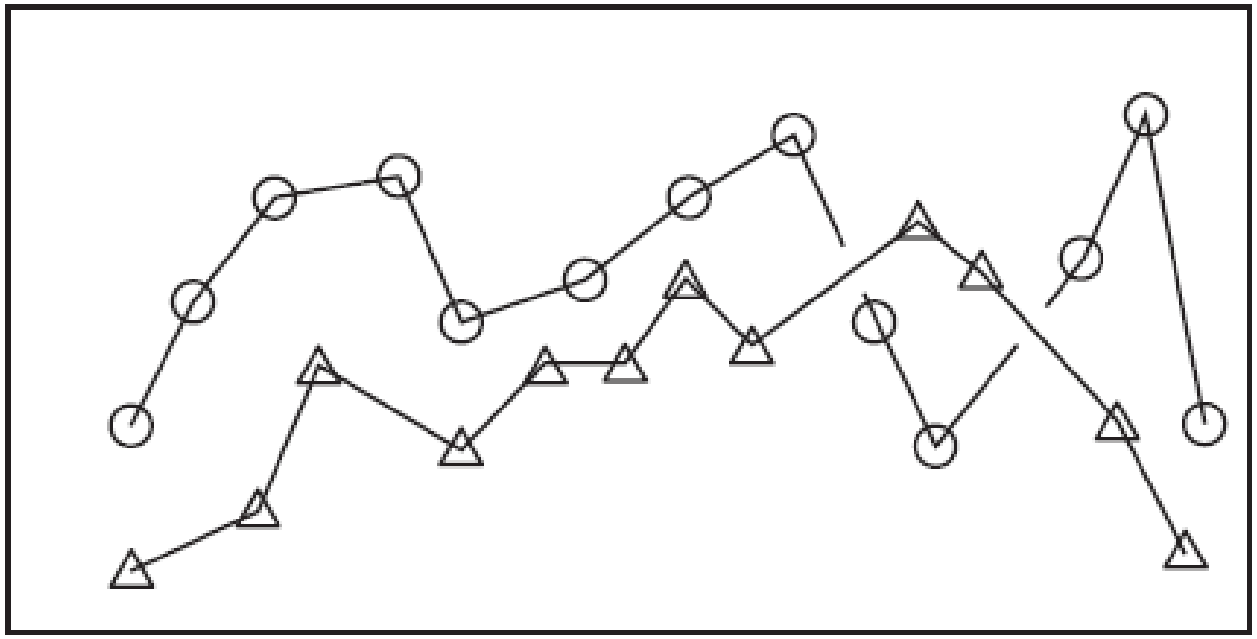
Regular Division of the Plane (M. C. Escher)



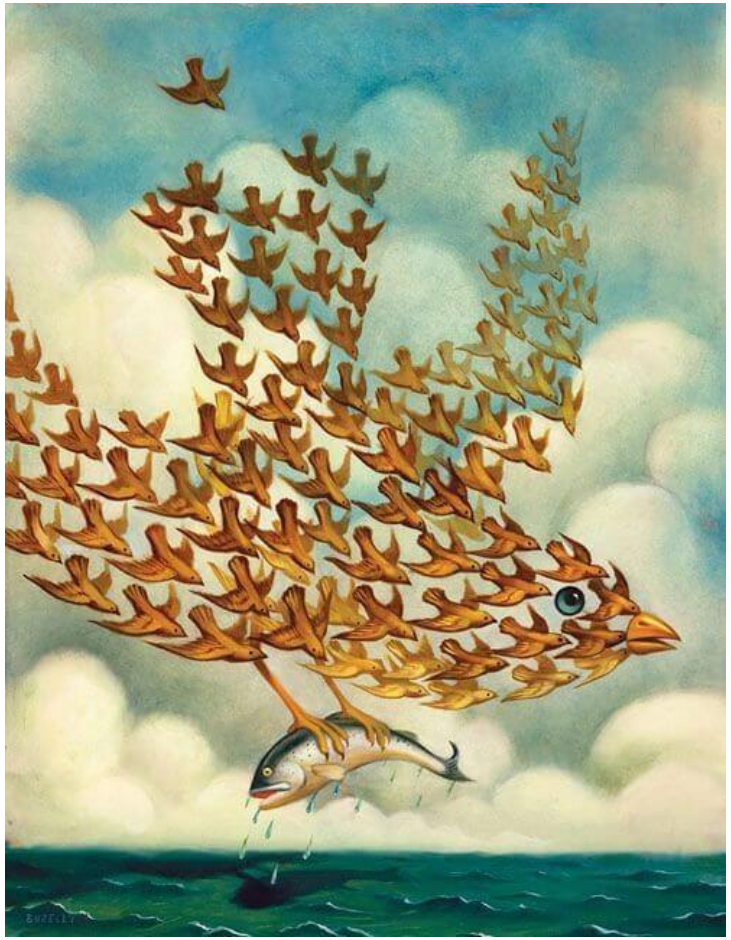
Fonte

Destino comum

- Elementos que compartilham uma direção ou movimento são percebidos como uma unidade



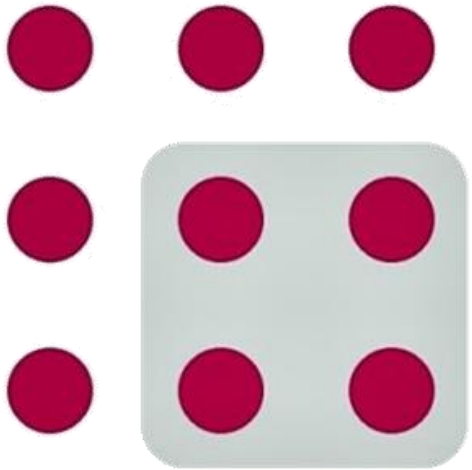
Adaptado de:
HEALY, K. **Data Visualization: A Practical Introduction.** Princeton University Press, 2019.

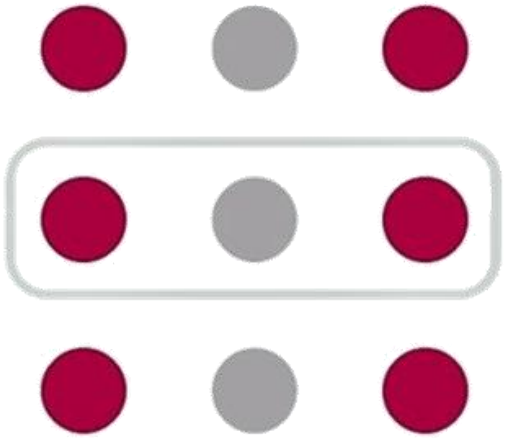


Fonte .

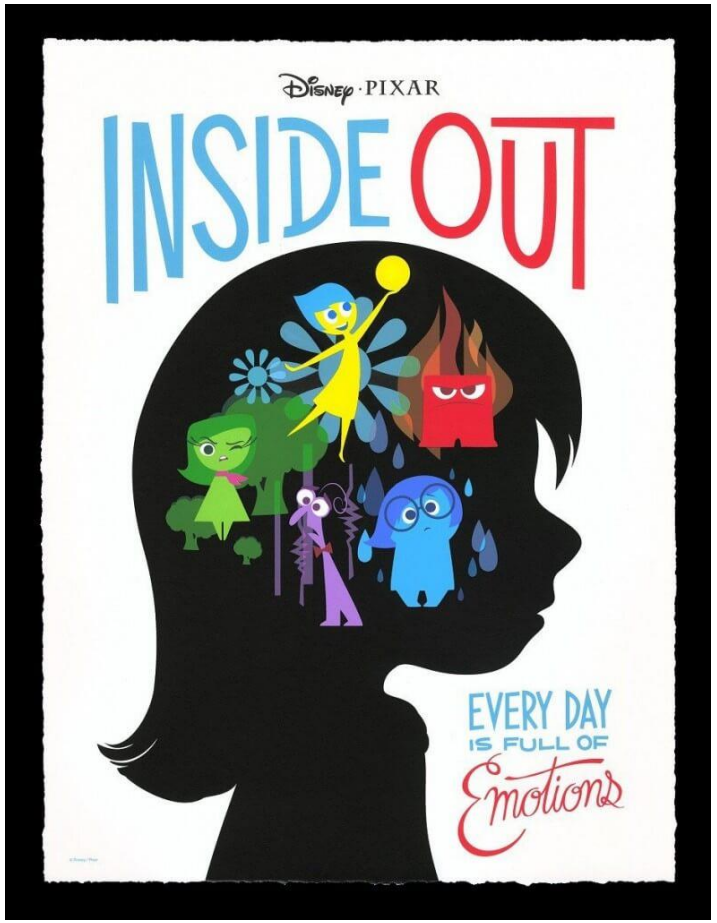
Região comum

- Elementos dentro de uma mesma área fechada são associados como pertencentes a um mesmo grupo





Fonte



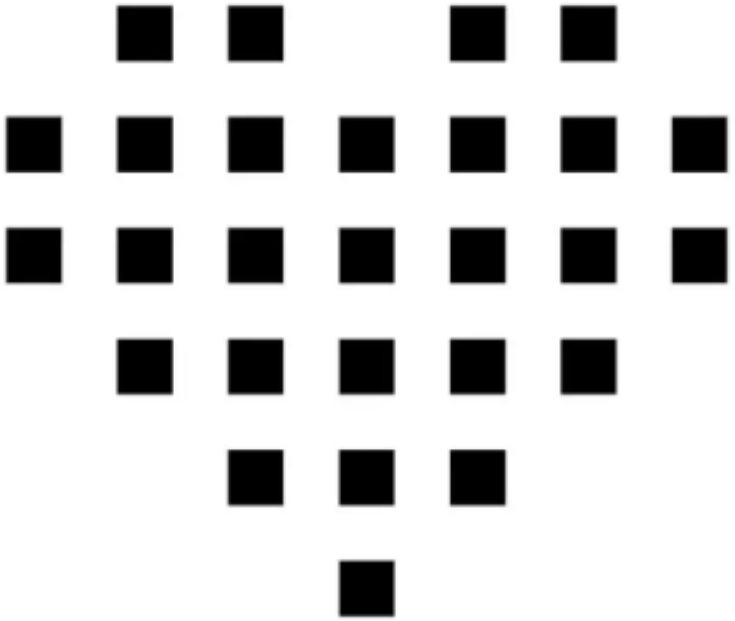
Segregação

- Capacidade perceptiva de separar ou destacar unidades do todo em função de diferenças de tamanho, forma, cor, espessura

Fonte

Unificação

- Tendência a visualizar elementos com semelhança, proximidade, fechamento e continuidade e simetria como uma única forma



Fonte

Fonte

Decodificando figuras visualmente

- A decodificação e interpretação de figuras é fruto da combinação entre:
  - Percepção visual (Limites, contrastes, cores)
  - Inferências sobre relações entre diferentes elementos visuais (Gestalt)

- Além destes componentes, há a necessidade de entendimento do que cada tipo de gráfico representa
  - O que é uma variável? Qual a diferença entre variável dependente e independente?
  - O que são eixos? Qual é o eixo usual para a variável independente?

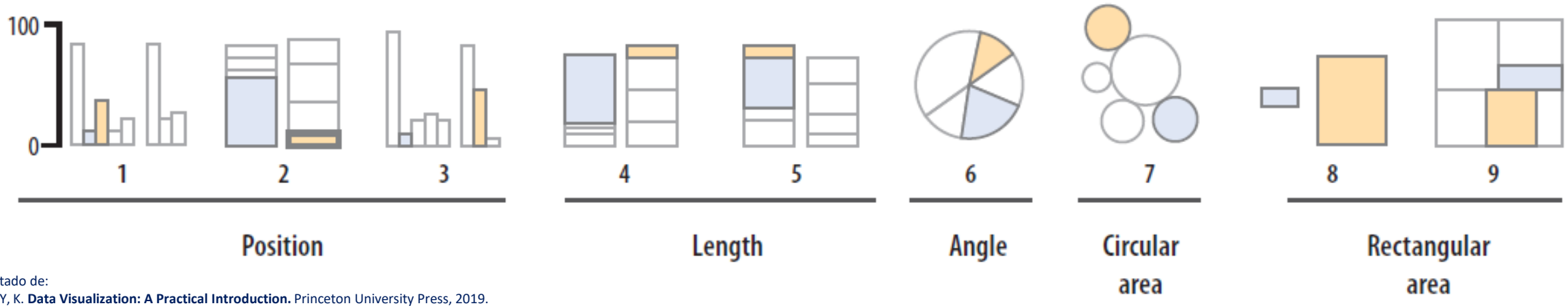
"A scatterplot is a visual representation of data, not a way to magically transmit pure understanding."

(K. Healy)

Há gráficos que são mais fáceis de decodificar do que outros...

- Experimentos de William Cleveland & Robert McGill (1984, 1987), Heer & Bostock (2010)

- Participantes deveriam estimar valores em um gráfico ou comparar valores entre gráficos distintos
  - Indicar a região menor e uma rápida estimativa da diferença entre a região maior e menor em termos de porcentagem



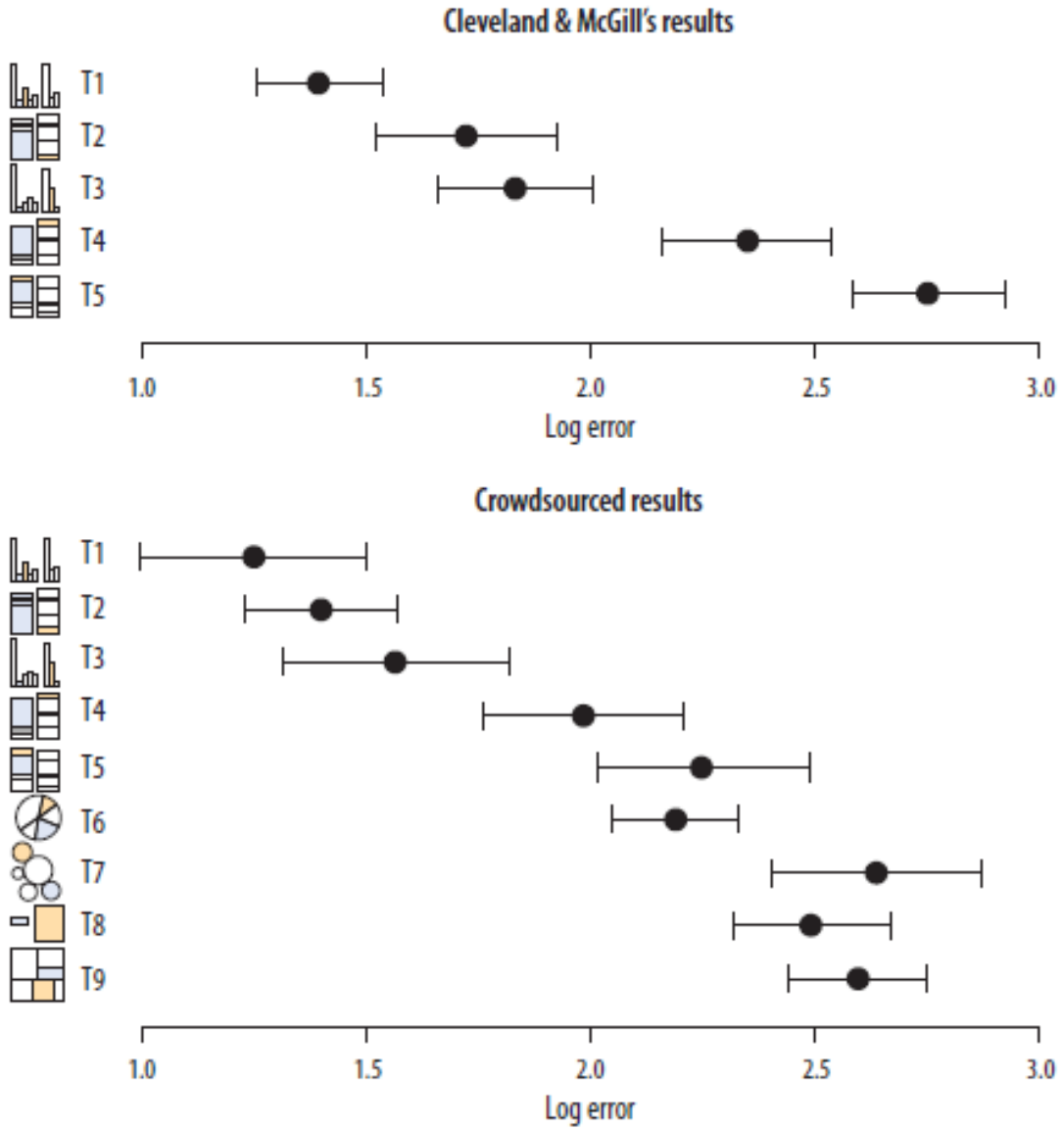
Adaptado de:
HEALY, K. **Data Visualization: A Practical Introduction.** Princeton University Press, 2019.

Há gráficos que são mais fáceis de decodificar do que outros...

- Melhor performance para comparações usando uma linha de base comum

- Performance intermediária para comprimento e ângulos
  - Ângulos agudos e ângulos obtusos são sub- e superestimados, respectivamente

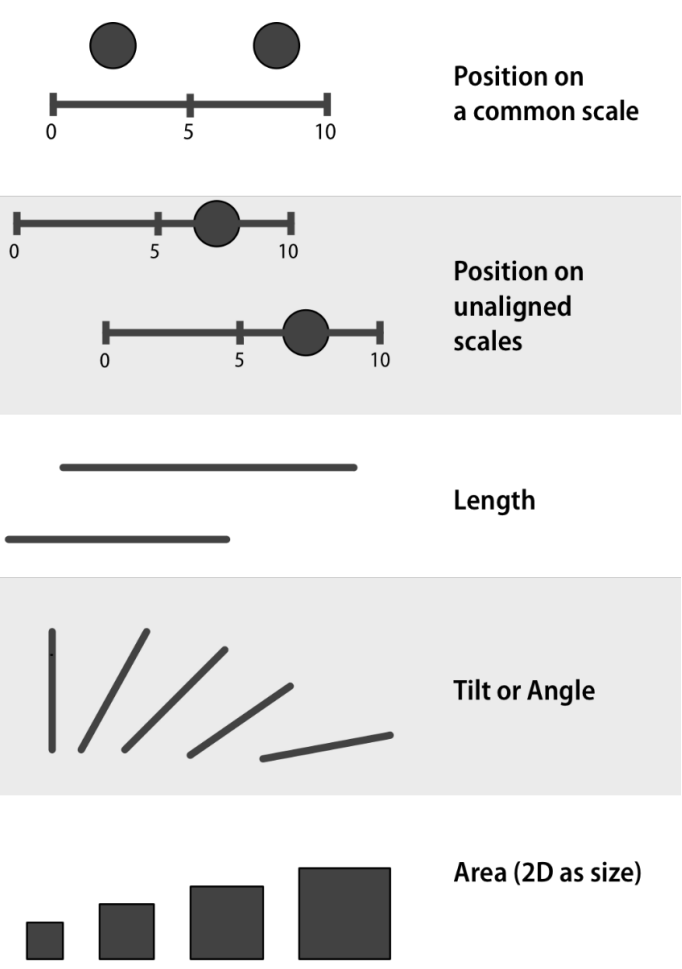
- Pior performance para áreas (incluindo áreas circulares)



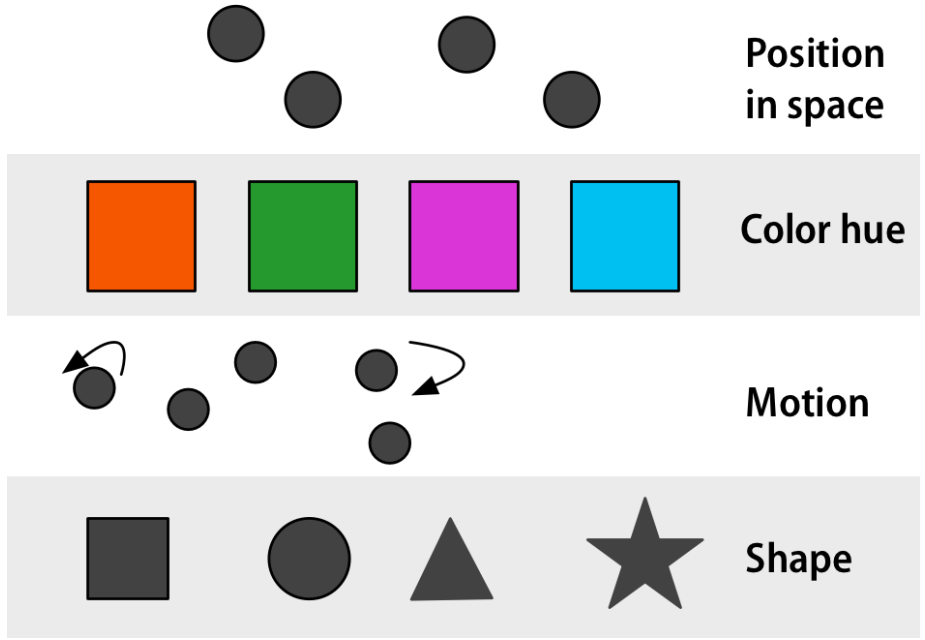
Adaptado de:
HEALY, K. **Data Visualization: A Practical Introduction.** Princeton University Press, 2019.

Há canais com melhor desempenho que outros...

- Variáveis ordinais



- Variáveis nominais



Adaptado de:
HEALY, K. **Data Visualization: A Practical Introduction.** Princeton University Press, 2019.

Sobre o uso de softwares

- Se há uma “constante” na área de visualização de dados, é a mudança nos métodos e softwares para elaboração de figuras ao longo do tempo

- Importância de focar em aprender visualização de dados a partir dos princípios gerais, e não a partir de um software específico

- Embora não seja o foco, é importante mencionar que existem softwares melhores que outros para essa finalidade

Sobre o uso de softwares

- Automatização sempre que possível:
  - Figuras devem preferencialmente serem geradas automaticamente como parte do processo de análise dos dados
  - Devem ser geradas o mais próximo possível da versão final, sem necessidade de pós-processamento manual

- Por que evitar processamento manual?
  - Ao realizar edições manuais, sua figura deixa de ser reproduzível por outras pessoas
  - Quanto maior a quantidade de edições, maior a procrastinação quando é necessário refazer a figura
  - Possibilidade de esquecimento das etapas e parâmetros utilizados durante a edição, impossibilitando a obtenção de uma nova figura que seja similar à original

Sobre o uso de softwares

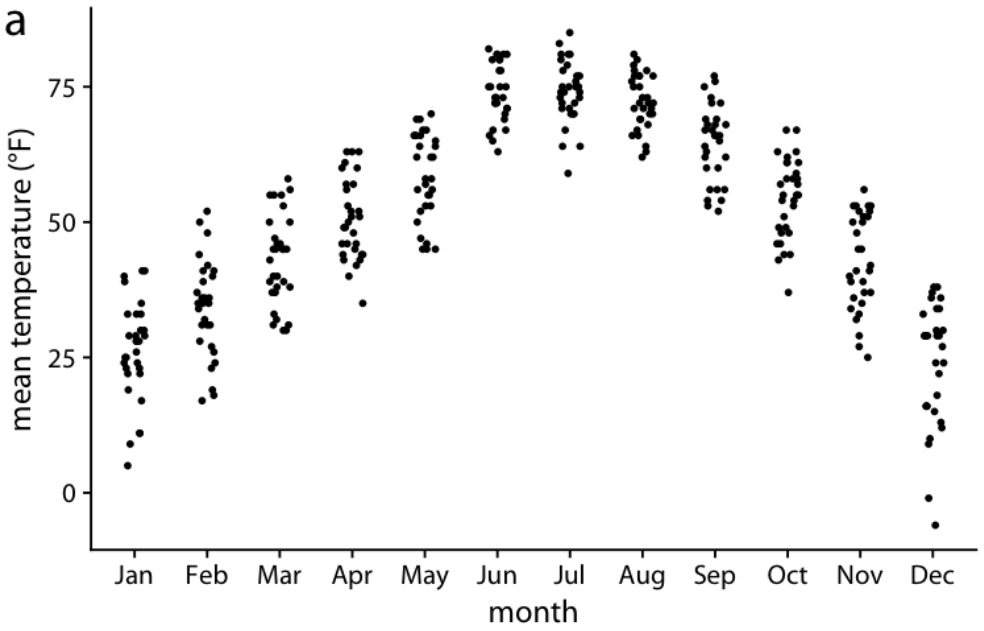
- Abolir o uso de softwares interativos:
  - Sim, Excel, estou olhando para você :)
  - Softwares interativos, em sua essência, fazem com que todo o processo de elaboração da figura seja manual (e não somente alguns eventuais ajustes)
  - Reprodutibilidade do processo ainda mais comprometida

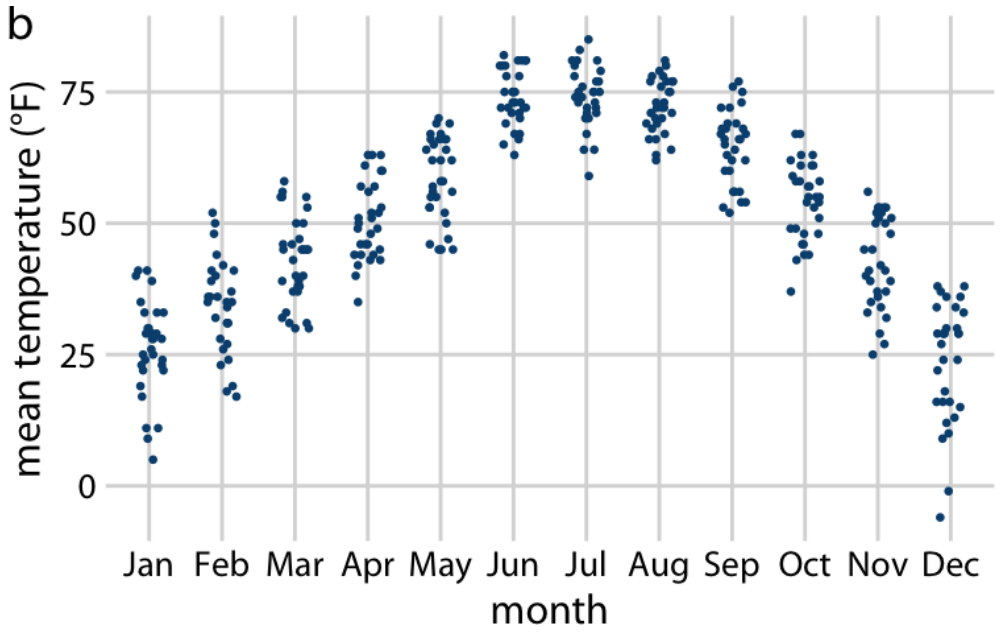
- Então, que software utilizar?

O melhor software de visualização de dados é aquele que funciona para você, e permite com que você faça as figuras que precisa...

Desde que atendidas algumas necessidades básicas:

- Reprodutibilidade e repetibilidade
  - Reprodutibilidade: capacidade de recriar uma figura similar a partir dos dados brutos, quando o método de criar o gráfico e as transformações de dados estão especificadas
  - Repetibilidade: capacidade de criar uma figura idêntica (em nível de pixels) a partir dos dados brutos, de modo que até mesmo elementos aleatórios (*jitter* / espalhamento de pontos) sejam idênticos





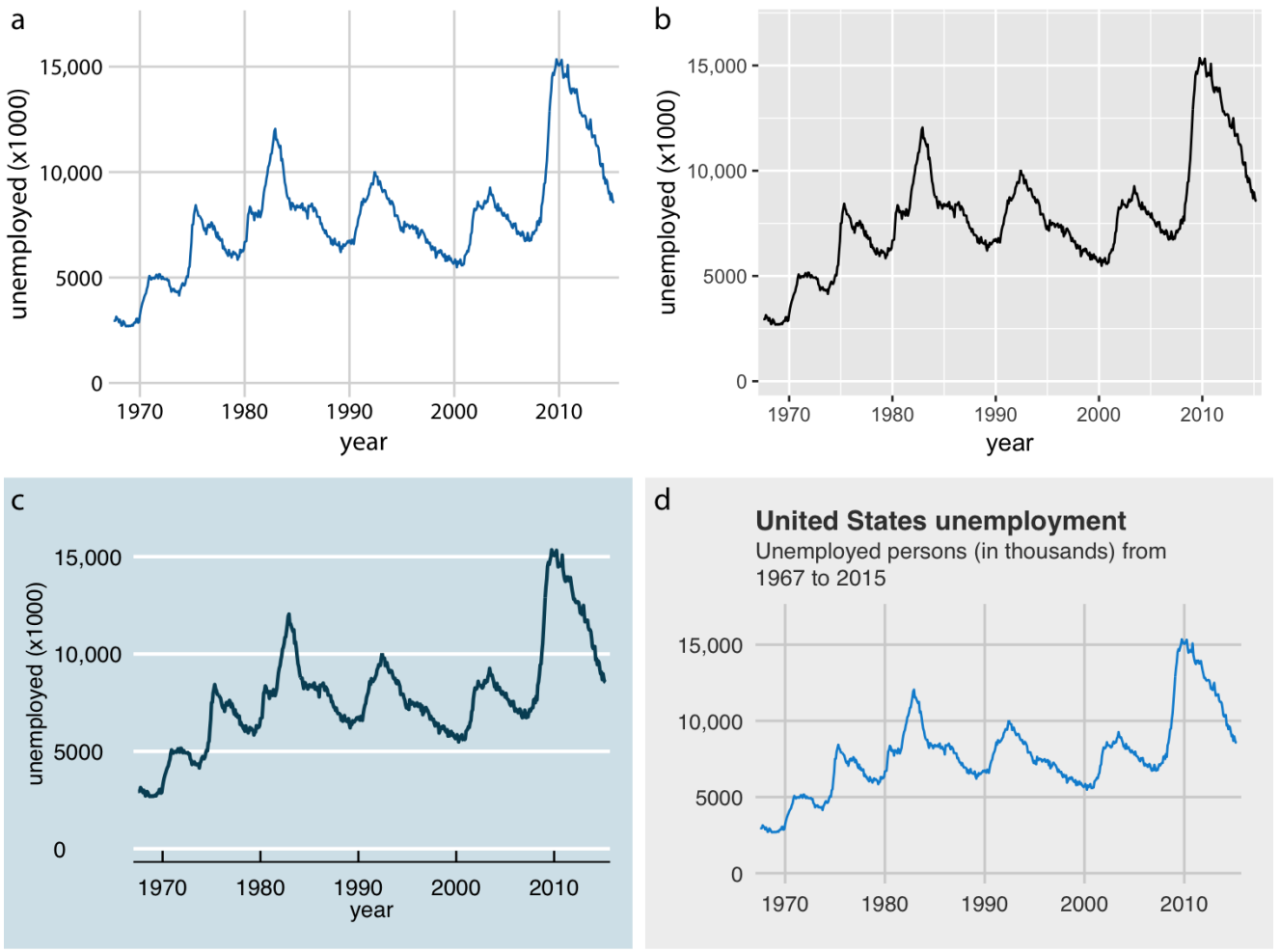
Adaptado de:
WILKE, C. O. **Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures.** O'Reilly Media, 2019.

Desde que atendidas algumas necessidades básicas:

- Exploração vs. apresentação dos dados
  - Nem sempre sabemos qual tipo de gráfico será utilizado já no começo do processo
  - Possibilidade de teste de vários tipos de gráfico sem que sejam necessários ajustes trabalhosos nos dados brutos (ex: reformatação inteira de uma tabela)
  - Após a exploração, o software deve permitir que os dados possam ser apresentados em figuras bem elaboradas, sem que ajustes manuais excessivos sejam necessários

Desde que atendidas algumas necessidades básicas:

- Separação entre conteúdo e design
  - O software deve possibilitar que conteúdo e design sejam ajustados de forma separada
  - Editar a parte visual e estética do gráfico sem ter que editar os dados brutos (e vice-versa)



Adaptado de:
WILKE, C. O. **Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures.** O'Reilly Media, 2019.