Introdução a Ciência de Dado

Professor: Alex Pereira

Escopo desta Disciplina

- Story Telling com dados
  - Teoria, e
  - Prática
    - ✓ Com Seaborn e D3/Notebooks Observable
- Construção de Dashboards (Paineis)
  - Teoria, e
  - Prática
    - ✓ Com Google Data Studio
- Coleta de Dados
  - Webscrapying
- Manipulação e Validação de Dados
  - Com pandas
- Armazenamento dos dados no BigQuery (Google)

A escolha das Ferramentas, Aplicações e trade-offs

- Custo de Transação, é o custo para
  - **realizar** qualquer negociação econômica ao participar de um mecado
    - ✓ custo/tempo de planejamento, decisão, mudança de planos, resolução de disputas, transporte/entrega/forma de consumo, pós-venda, taxas de pagamento, comissões
- Tendências do mercado de TI
  - Zerar o custo de licença e diminuir os custos de transação da adesão
    - ✓ Gmail, Google Drive, Google Analytics, Google Data Studio, Redes Sociais
      - Extrair lucro (propaganda) a partir do acesso aos dados
        - Quando a licença é gratuita o produto é você.
  - Free-tier (Amostra grátis)
    - ✓ Google Cloud Platform (Bigquery e outros), AWS (Redshift e outros)
- Consequências para a Administração Pública
  - Celeridade e economicidade
    - ✓ Ao custo de compartilhar os dados da administração pública

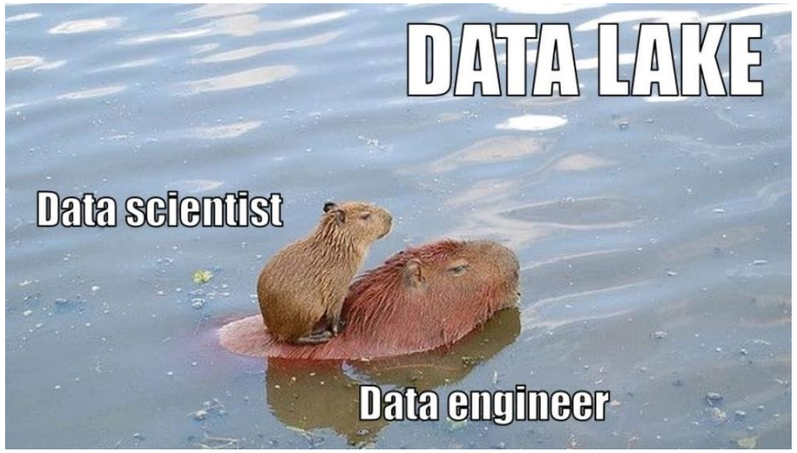
As ferramentas deste curso

- Google Data Studio
  - Sem custo de licença
  - Integração facilitada do com o BigQuery, Google Sheets e Analytics
  - Infraestrutura gerenciada pelo Google
    - ✓ Única solução de Dashboard as a Service (DaaS)
      - sem custo de licença em que se pode publicar abertamente um painel
  - Baixo custo de transação de adesão
- Notebooks Observable (para contar uma história com dados)
  - Sem custo de licença para publicar e reusar
    - ✓ Os gráficos, visualizações e notebooks
  - Infraestrutura gerenciada pela Empresa
  - Baixo custo de transação de adesão

Contar uma história com dados

- Habilidade fundamental, de importância crescente
  - + dados digitais
  - Potencializa sua capacidade de
    - ✓ argumentação e convencimento
- Formação de agenda de políticas públicas
  - Teoria da Lata do Lixo (Garbage Can)
    - ✓ Uma coleção de escolhas encontra um problema
      - Entidades envolvidas: soluções, problemas e tomadores de decisão
    - ✓ Nesse contexto, **publicar** suas soluções na forma de **histórias com dados**
      - Pode aumentar a probabilidade de soluções encontrarem os problemas
        - Ferramentas com baixo custo de transação proporcionam esse benefício
  - O papel do cientista de dados
    - ✓ Descobrir e informar

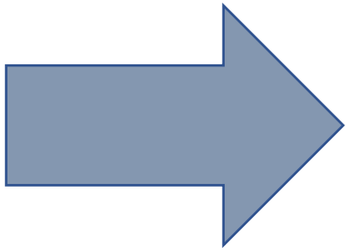
Engenharia de Dados: poderia ser parte do escopo, mas ficou de fora ☹

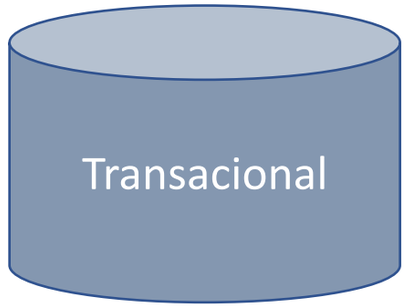


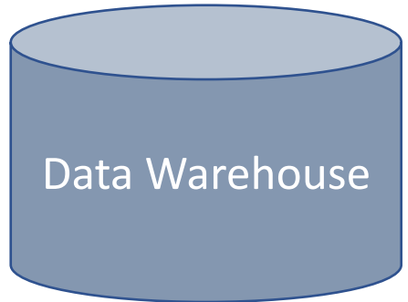
- Mover dados atendendo os requisitos funcionais e não funcionais
  - Inclusive a atualização frequente e automática.
  - De diversas fontes e formatos
    - ✓ Para diversos destinos e formatos

- Banco de Dados
- FTP
- API
- Google Sheet
- Drive/S3/Cloud
- Web/HTML
- ...

- json
- xml
- html
- Avro
- csv
- pdf
- doc
- ...

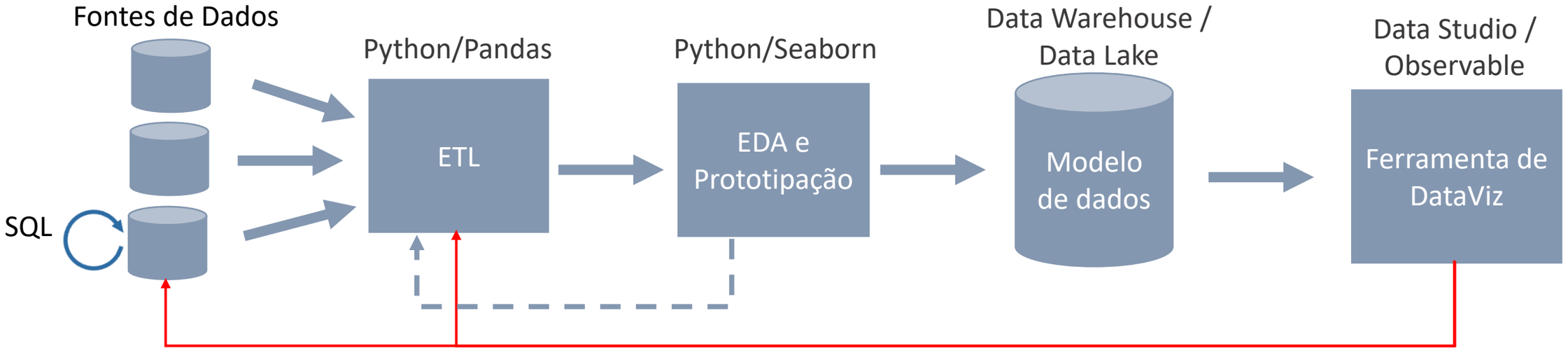






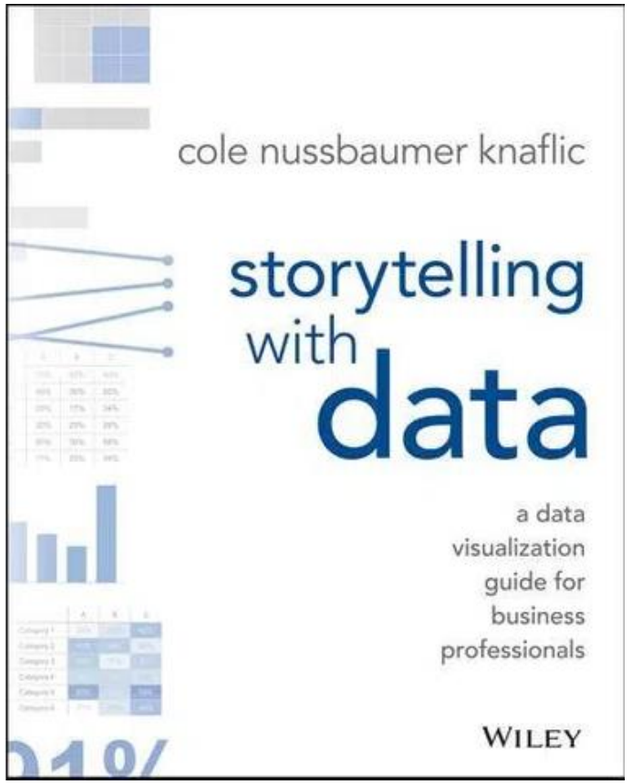
Sugestão de Metodologia de Trabalho

- Para elaborar painéis ou histórias com dados
  - Defina um problema, estude as variáveis disponíveis e
    - ✓ aplique uma técnica de ideação / Story Boarding
- Use uma metodologia iterativa de construção do modelo de dados
  - Minimize o tamanho das iterações
    - ✓ Iterações são inevitáveis e esperadas, **minimize o custo para iterar!**
      - Erro comum: validar as transformações somente na última etapa.



Storytelling with Data

- Cole Nussbaumer Knaflic
  - Autora do livro Storytelling with data
  - Saiu do Google para se tornar consultora
    - ✓ de Storytelling
  - Empreendeu e criou vários produtos relacionados
    - ✓ Podcast
    - ✓ Audiolivro
    - ✓ Blog
    - ✓ Workshops
    - ✓ Makeovers



Principais Propósitos da Elaboração de Gráficos

- Análise Exploratória de Dados (Propósito Pesquisar)
  - Investigar os dados a fim de se chegar a conclusões relevantes
  - Como procurar pérolas em ostras
    - ✓ Abrir 100 ostras (testar 100 hipóteses) para encontrar 2 pérolas
- Análise Explanatória (Propósito Comunicar)
  - Facilitar a Comunicação de um achado relevante
    - ✓ É um erro apresentar a análise exploratória (apresentar as 100 ostras),
    - ✓ Enquanto se deveria apresentar a análise explanatória (as 2 pérolas)
      - use o tempo da audiência para comunicar informação

Conte os números 3

7563950684 73

6586630375 76

8603726586 02

8465891078 30

Fonte: KNAFLIC, C. N. (2018)

Conte os números 3

7563950684733

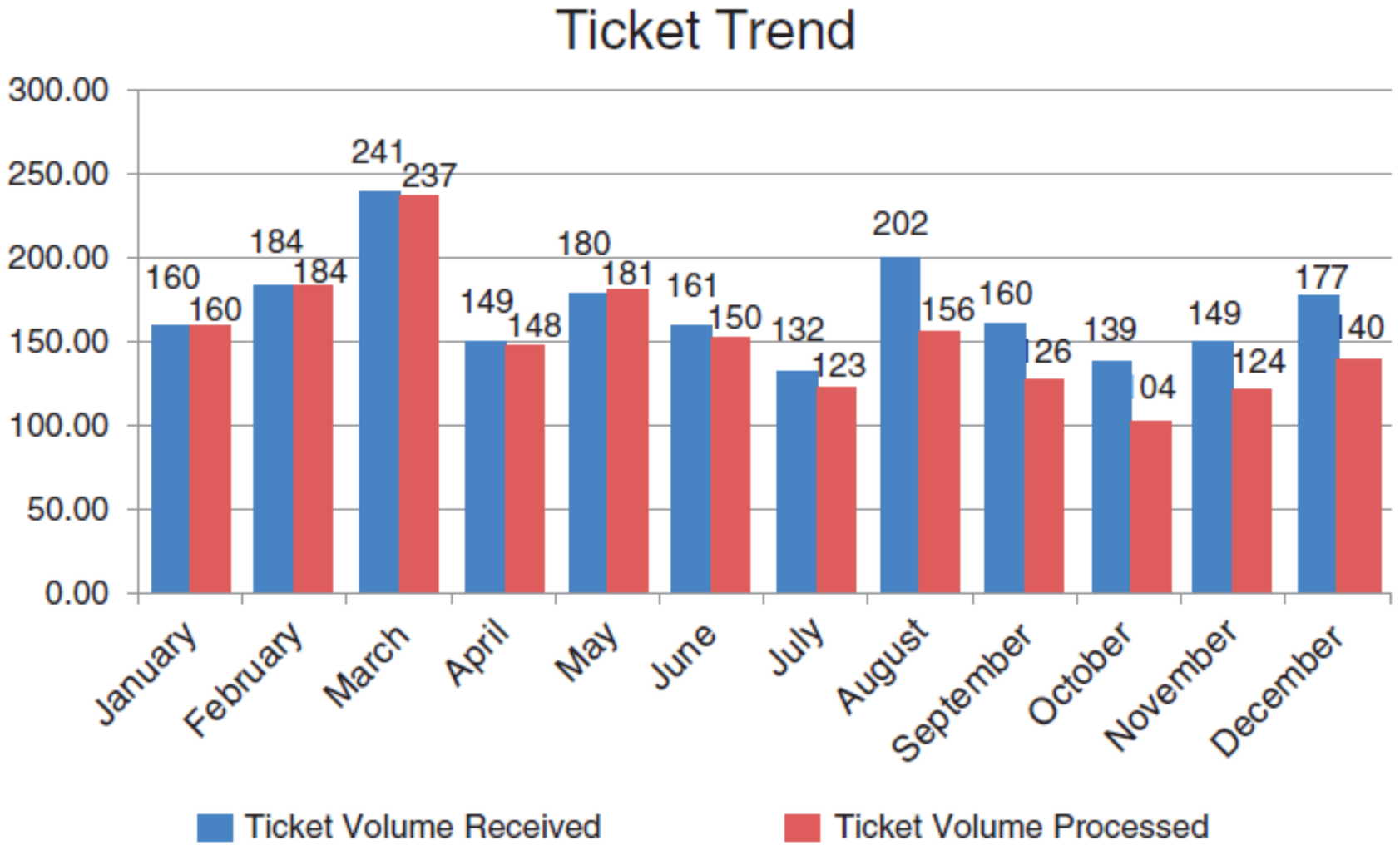
6586630375576

8603726586602

8465891078330

Fonte: KNAFLIC, C. N. (2018)

Exemplos e Contra-exemplos (1)



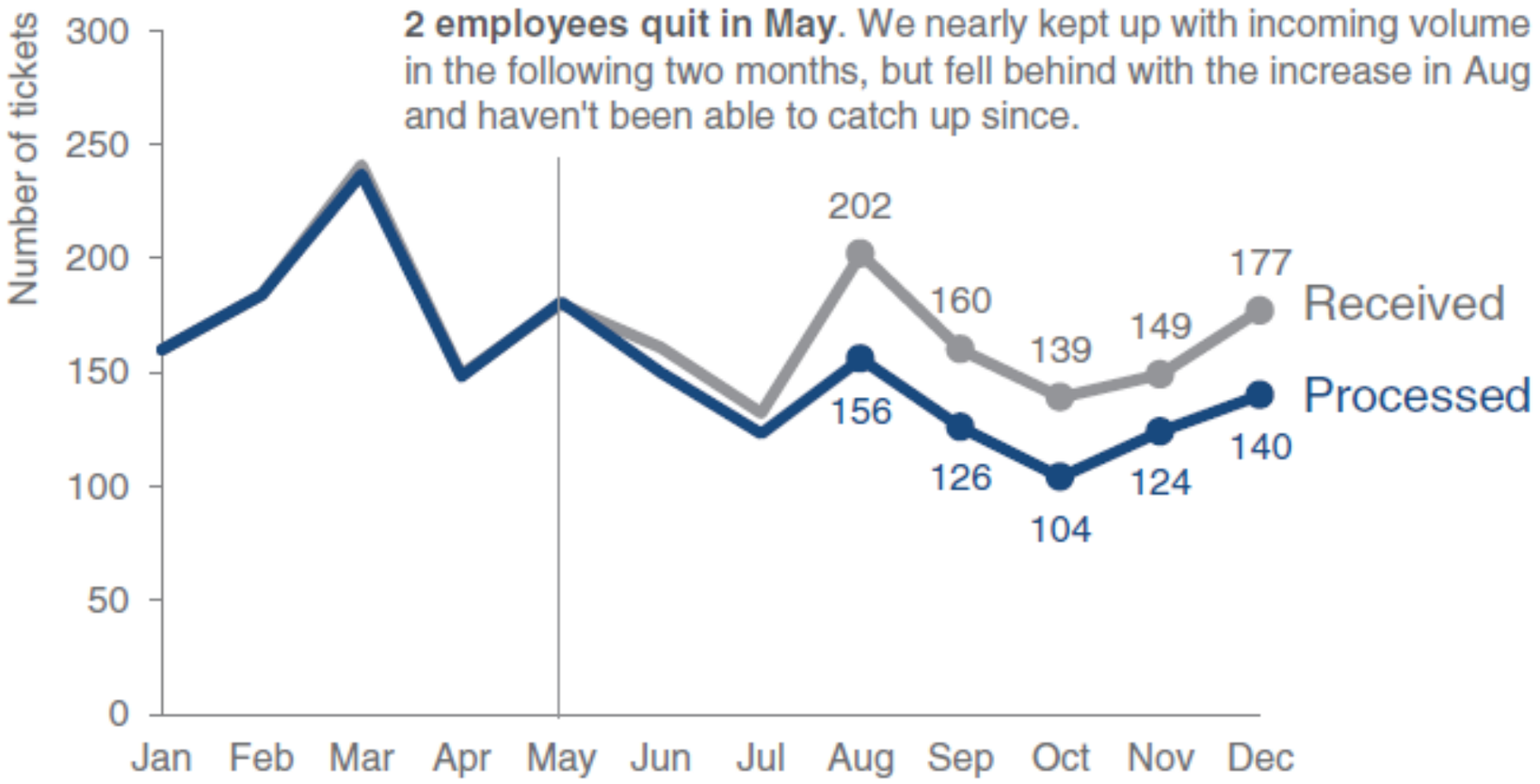
Fonte: KNAFLIC, C. N. (2018)

Exemplos e Contra-exemplos (2)

Please approve the hire of 2 FTEs

to backfill those who quit in the past year

Ticket volume over time



Fonte: KNAFLIC, C. N. (2018)

Atributos Pré-atenção

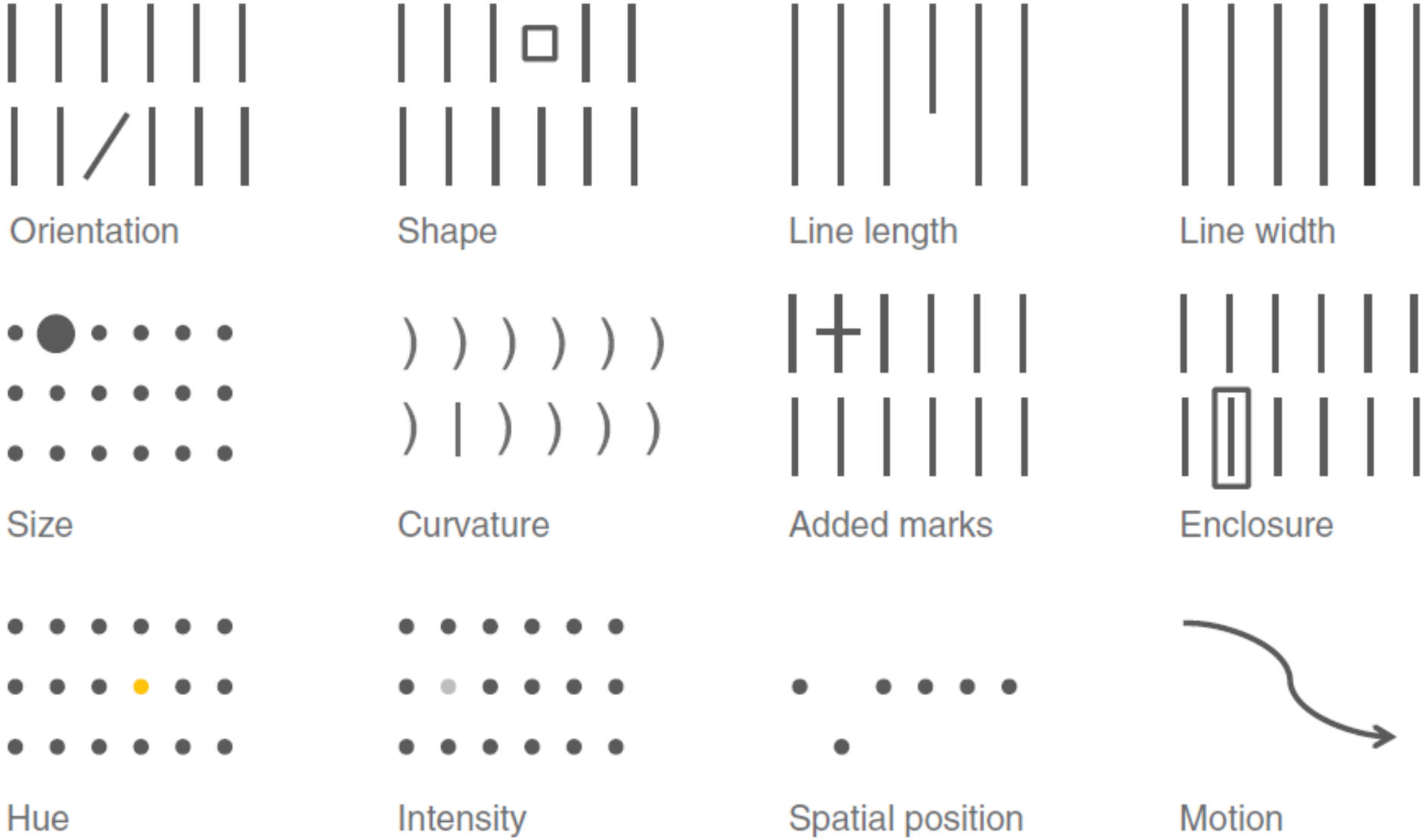


FIGURE 4.4 Preattentive attributes

Fonte: KNAFLIC, C. N. (2018)

Dois modos de Pensar (Daniel Kahneman)

- Rápido
- Intuitivo
- Não trabalhoso

17 x 24

- Lento
- Trabalhoso
- Deliberado
- Ordenado

Dando ouvidos a sua intuição

- Um bastão e uma bola custam R$ 1,10.
- O bastão custa um real a mais que a bola.

Quanto custa a bola?

- Mais de 50% dos estudantes de
  - Harvard, MIT e Princeton
    - ✓ deram uma resposta incorreta

Atributos Pré-atenção – Dois propósitos

- Direcionar a atenção do seu público
  - para onde você deseja que ele se concentre.
- Criar uma hierarquia visual de elementos para conduzir seu público
  - através das informações que você deseja comunicar
    - ✓ da maneira que deseja que eles as processem.

Atributos Pré-atenção – Exemplos em Texto

No preattentive attributes

What are we doing well? Great Products. These products are clearly the best in their class. Replacement parts are shipped when needed. You sent me gaskets without me having to ask. Problems are resolved promptly. Bev in the billing office was quick to resolve a billing issue I had. General customer service exceeds expectations. The account manager even called to check in after normal business hours.
You have a great company – keep up the good work!

Bold

What are we doing well? Great Products. These products are clearly the best in their class. Replacement parts are shipped when needed. You sent me gaskets without me having to ask. Problems are resolved promptly. Bev in the billing office was quick to resolve a billing issue I had. General customer service exceeds expectations. The account manager even called to check in after normal business hours.
You have a great company – keep up the good work!

Color

What are we doing well? Great Products. These products are clearly the best in their class. Replacement parts are shipped when needed. You sent me gaskets without me having to ask. Problems are resolved promptly. Bev in the billing office was quick to resolve a billing issue I had. General

Italics

What are we doing well? Great Products. These products are clearly the best in their class. *Replacement parts are shipped when needed*. You sent me gaskets without me having to ask. Problems are resolved promptly. Bev in the billing office was quick to resolve a billing issue I had. General

Fonte: KNAFLIC, C. N. (2018)

Atributos Pré-atenção – Hierarquia

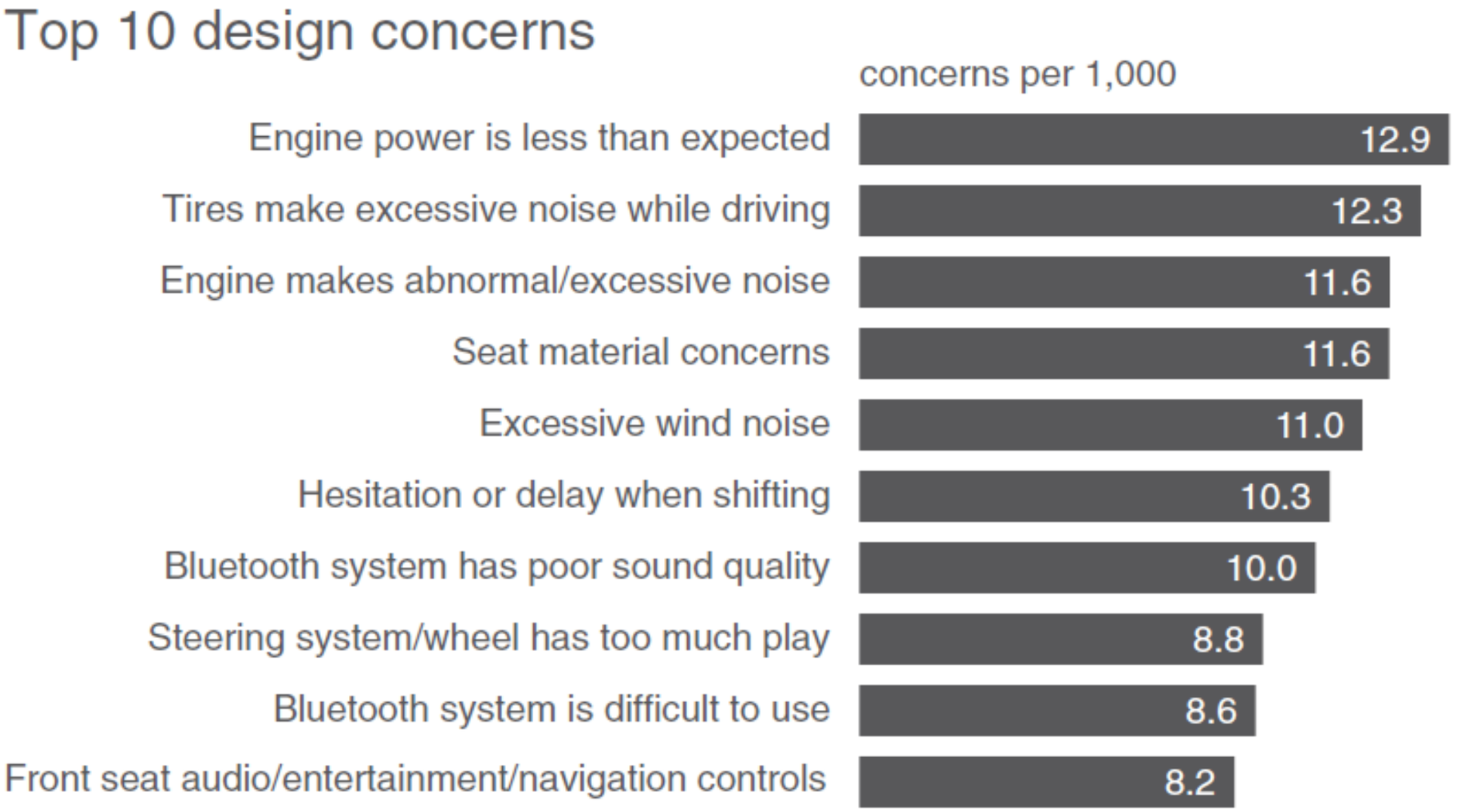
What are we doing well?

Themes & example comments

- **Great products**: "These products are clearly the best in class."
- **Replacement parts are shipped when needed**: "You sent me gaskets without me having to ask, and I really needed them, too!"
- **Problems are resolved promptly**: "Bev in the billing office was quick to resolve a billing issue I had."
- **General customer service exceeds expectations**: "The account manager even called after normal business hours. *You have a great company - keep up the good work!*"

Fonte: KNAFLIC, C. N. (2018)

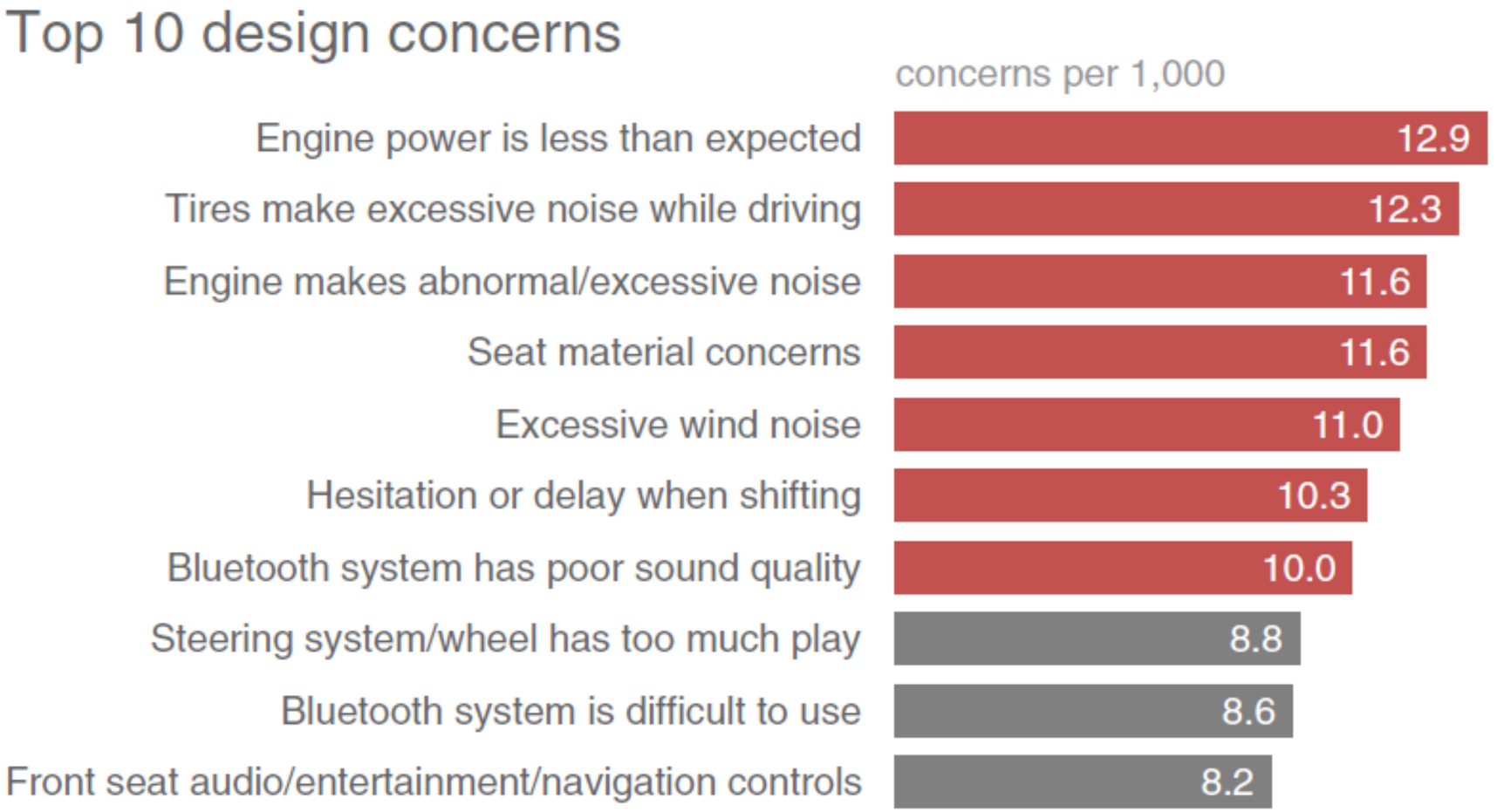
Gráfico sem o uso de Atributos Pré-atenção



Fonte: KNAFLIC, C. N. (2018)

Gráfico com o uso de Atributos Pré-atenção

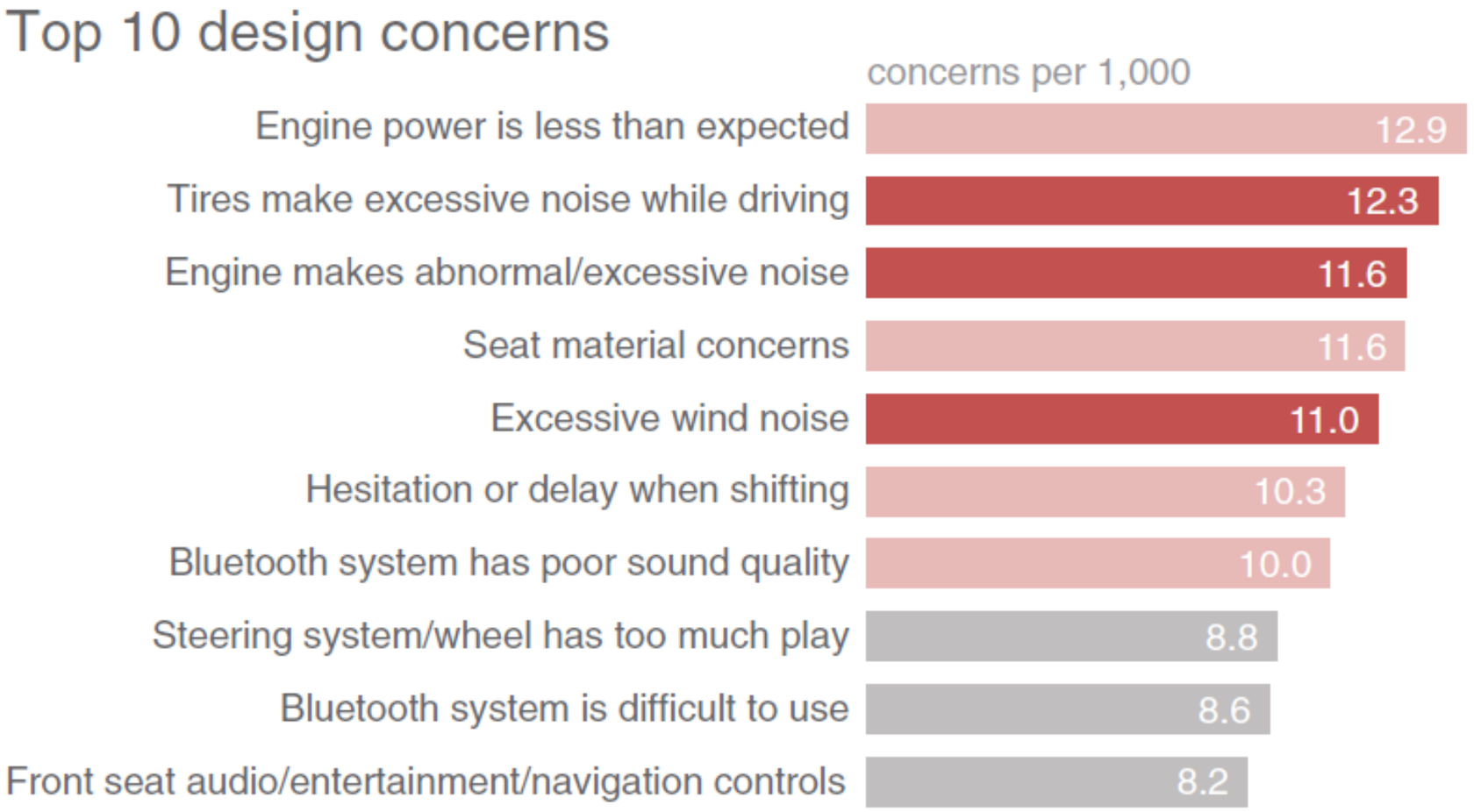
7 of the top 10 design concerns have 10 or more concerns per 1,000.
Discussion: is this an acceptable default rate?



Fonte: KNAFLIC, C. N. (2018)

Mais conclusões expostas com ajuda dos atributos

Of the top design concerns, three are noise-related.



Comments indicate that **noisy tire issues** are most apparent **in the rain**.

Complaints about **engine noise** commonly cited **after the car had not been driven for a while**.

Excessive **wind noise** is noted primarily in **freeway driving at high speeds**.

Fonte: KNAFLIC, C. N. (2018)

Clutter (Desordem/tumulto) = Inimigo

- Cada elemento adicionado a uma página (tela)
  - demanda esforço cognitivo do seu público
- Identifique qualquer coisa que não adiciona informação valiosa
  - e elimine-a.
- Esforço cognitivo é demandado para assimilar novas informações
  - Como designer de informação, como minimizar o esforço demandado da audiência ?
- Data-ink ratio (razão dados-tinta)
  - Quanto maior a proporção de tinta gastos em dados, melhor
- Signal-noise ratio (razão sinal-ruído)
  - Quanto mais sinal pra mesma quantidade de ruído, melhor

Princípios Gestalt de Percepção Visual

- Definidos pela Escola Gestalt de Psicologia nos anos 1900
  - Para entender como indivíduos percebem ordem no mundo a sua volta
    - ✓ por meio da absorção de estímulos visuais
- 6 princípios
  - proximity,
  - similarity,
  - enclosure,
  - closure,
  - continuity, e
  - connection.

Proximity (Proximidade)

- Tendemos a perceber objetos fisicamente próximos
  - como pertencentes ao mesmo grupo

Fonte: KNAFLIC, C. N. (2018)

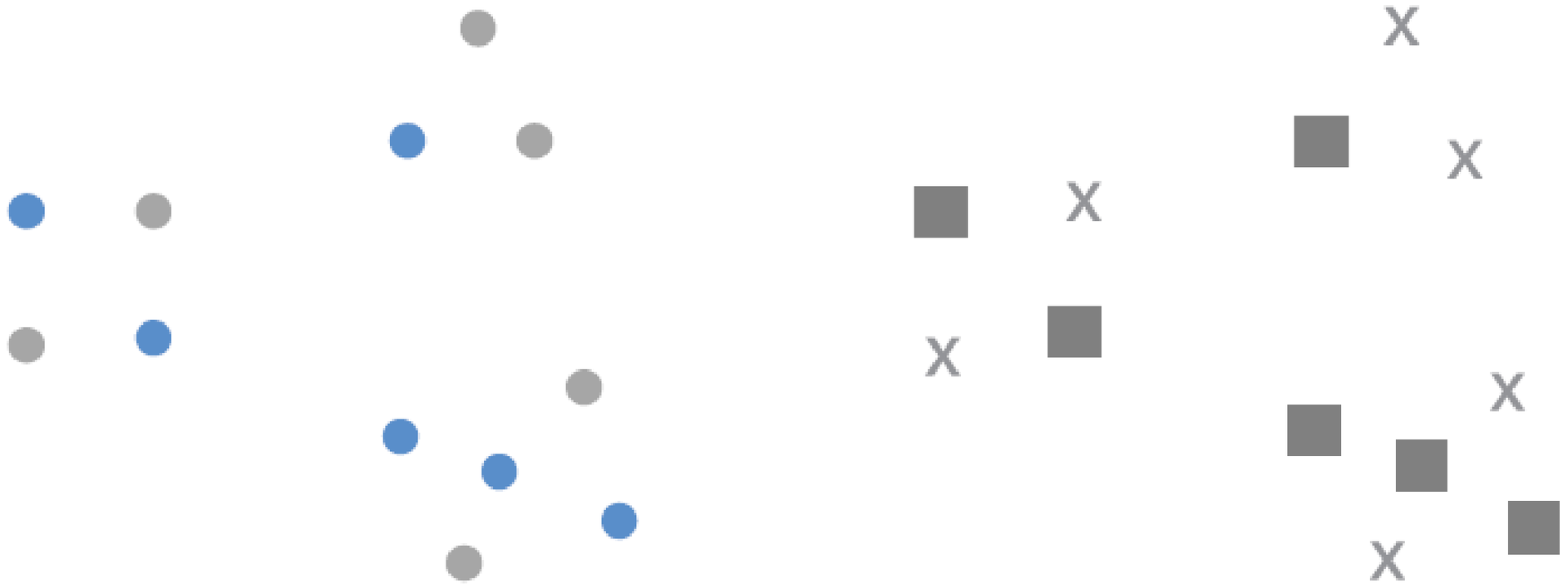
Proximity (Proximidade)

- Ao mudar o espaçamento entre os pontos
  - nossos olhos são atraídos para as colunas (à esquerda)
    - ✓ ou para as linhas (à direita)

Fonte: KNAFLIC, C. N. (2018)

Similarity (Similaridade)

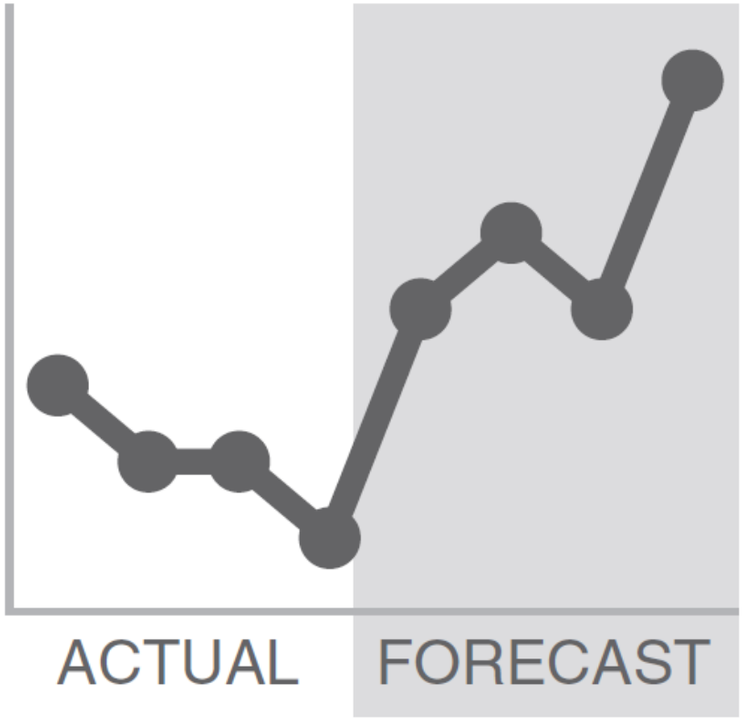
- Objetos de cor, forma, tamanho ou orientação similar
  - são percebidos como pertencentes a um mesmo grupo



Fonte: KNAFLIC, C. N. (2018)

Enclosure (Enclausuramento)

- Objetos fisicamente enclausurados
  - são percebidos como pertencentes a um mesmo grupo



Fonte: KNAFLIC, C. N. (2018)

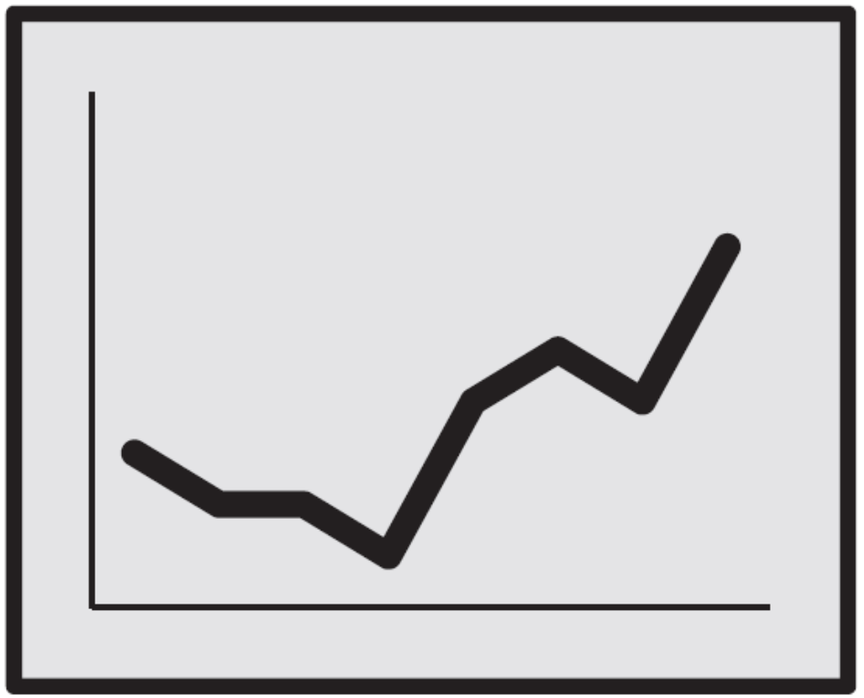
Closure (Fechamento)

- Pessoas gostam de encaixar construtos em coisas que já conhecem
  - esses tipos de objetos são percebidos como apenas 1.

Fonte: KNAFLIC, C. N. (2018)

Closure (Fechamento)

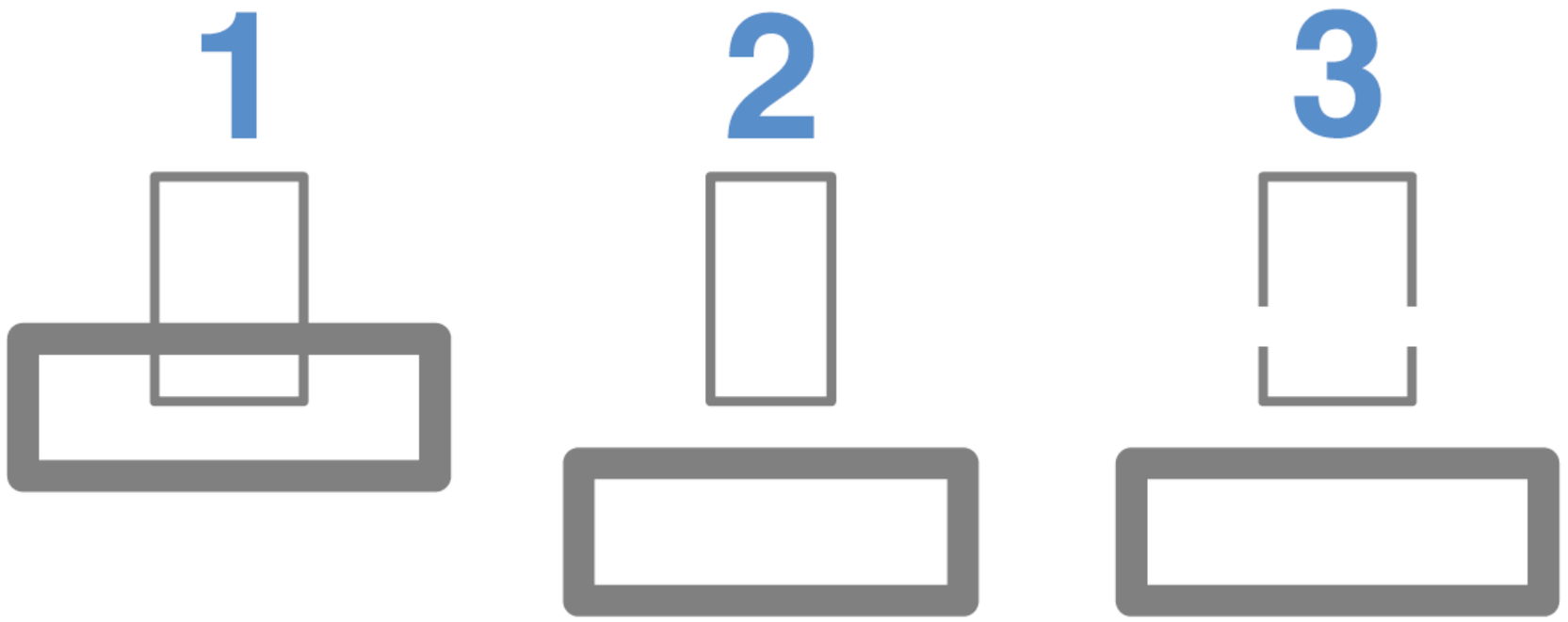
- Pelo princípio do fechamento
  - as bordas externas são desnecessárias



Fonte: KNAFLIC, C. N. (2018)

Continuity (Continuidade)

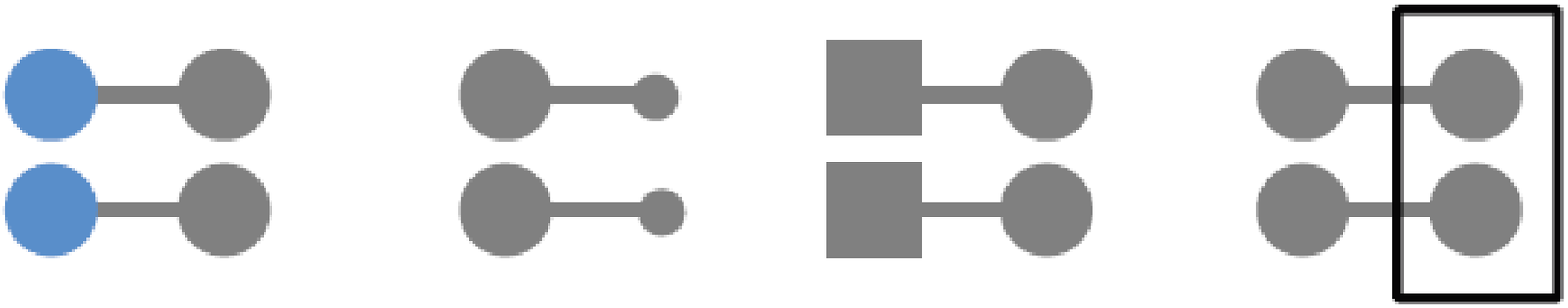
- Nossos olhos procuram continuidade, até onde não existe.
- Se separarmos os objetos em (1)
  - esperamos ver os objetos em (2)
    - ✓ Mas a configuração em (3) também é possível



Fonte: KNAFLIC, C. N. (2018)

Connectivity (Conexção)

- Percebemos objetos conectados como pertencentes ao mesmo grupo
  - a conexão tem um valor associado maior do que
    - ✓ cor, tamanho e forma similares



Fonte: KNAFLIC, C. N. (2018)

Connectivity (Conexão)

- Usamos este princípio no gráfico de linhas
  - para ajudar nossos olhos enxergar ordem nos dados

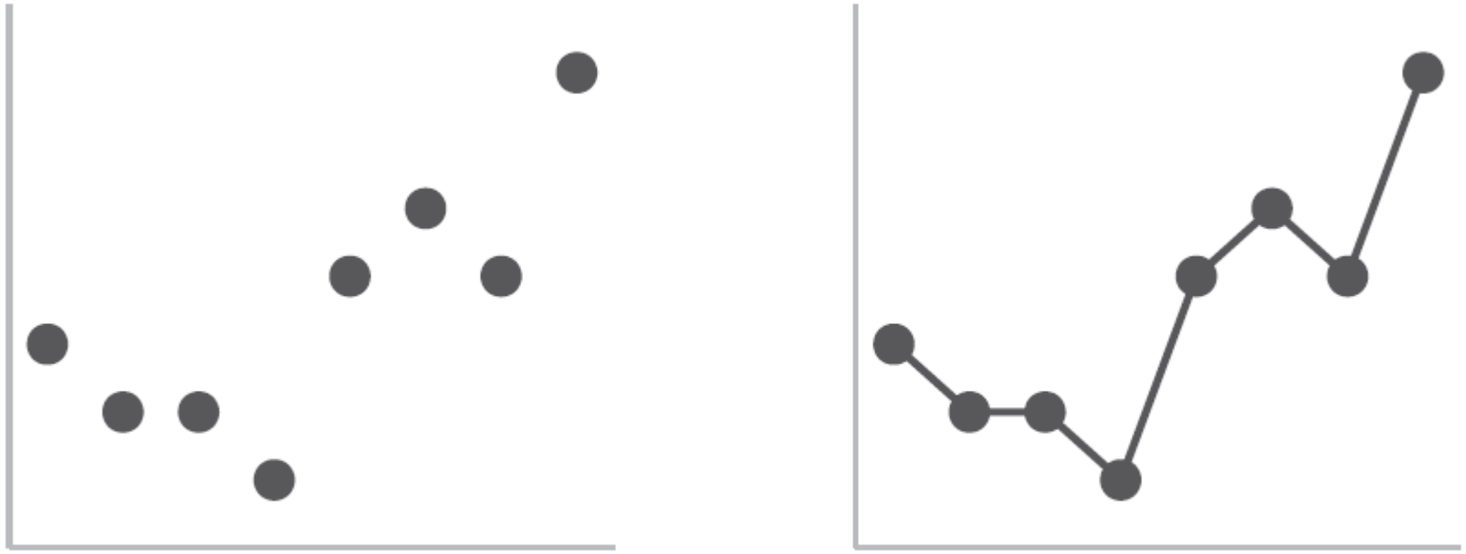
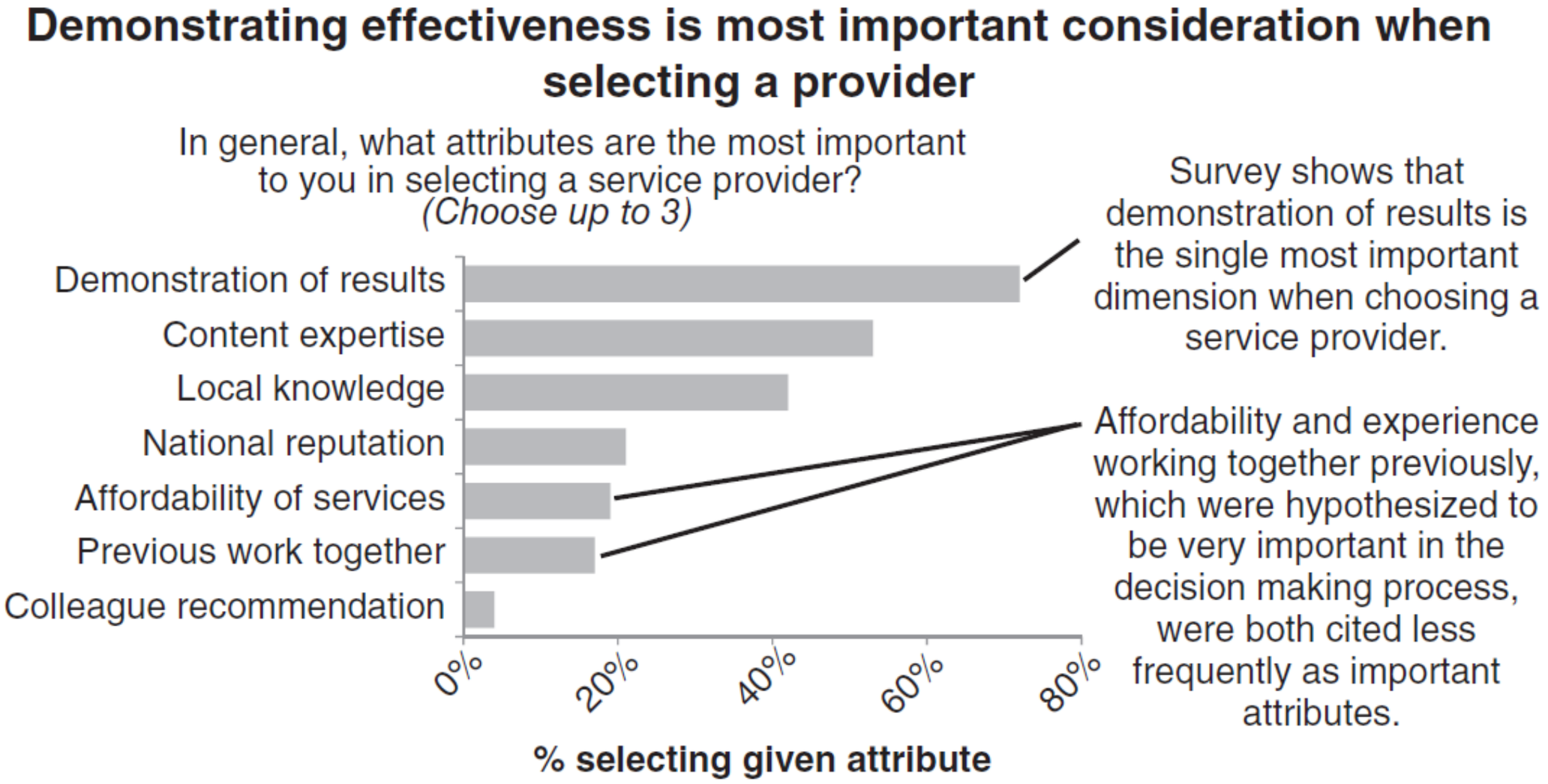


FIGURE 3.12 Lines connect the dots

Fonte: KNAFLIC, C. N. (2018)

Aplicando os princípios



Fonte: KNAFLIC, C. N. (2018)

Aplicando os princípios

Demonstrating effectiveness is most important consideration when selecting a provider

In general, **what attributes are the most important** to you in selecting a service provider?

(Choose up to 3)

Data source: xyz; includes N number of survey respondents.
Note that respondents were able to choose up to 3 options.

Survey shows that **demonstration of results** is the single most important dimension when choosing a service provider.

Affordability and **experience working together previously**, which were hypothesized to be very important in the decision making process, were both cited less frequently as important attributes.

Fonte: KNAFLIC, C. N. (2018)

Aplicando os princípios

Demonstrating effectiveness is most important consideration when selecting a provider

In general, **what attributes are the most important** to you in selecting a service provider?

(Choose up to 3)

Data source: xyz; includes N number of survey respondents.
Note that respondents were able to choose up to 3 options.

Principal alteração foi o alinhamento à esquerda, em vez de centralizado.

Survey shows that **demonstration of results** is the single most important dimension when choosing a service provider.

Affordability and **experience working together previously**, which were hypothesized to be very important in the decision making process, were both cited less frequently as important attributes.

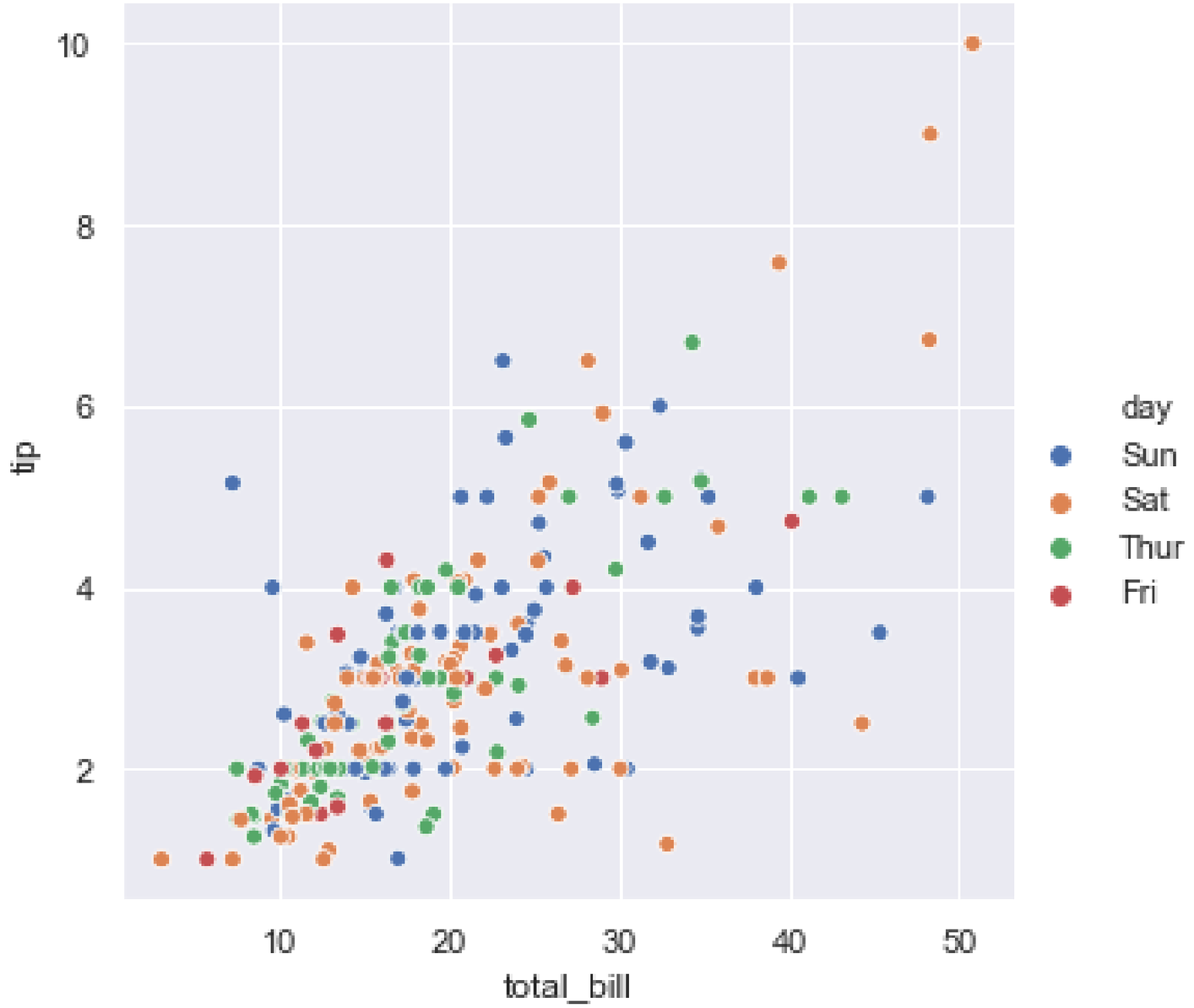
Fonte: KNAFLIC, C. N. (2018)

EDA e Prototipação com Seaborn

- Dataset tips (gorjetas)

	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.50	Male	No	Sun	Dinner	3
3	23.68	3.31	Male	No	Sun	Dinner	2
4	24.59	3.61	Female	No	Sun	Dinner	4

relplot de dispersão (scatter)



Seaborn (Revisão)

```
import seaborn as sns
sns.set() # Ajusta valor padrão de tema, escala e paleta de cor
tips = sns.load_dataset("tips") # Carrega um dataset de gorgetas
sns.relplot(x="total_bill", y="tip", hue="day", data=tips)
```

- `relplot` representa uma categoria de gráficos
  - Que mostram relação entre duas variáveis.
  - Atalho para os gráficos de dispersão (scatterplot) e gráfico de linha
    - ✓ kind='scatter' ou kind='line' equivale a scatterplot() ou lineplot()
  - `data=tips` – dataset a ser plotado
  - `x="total_bill"` – coluna plotada no eixo x
  - `y="tip"` – coluna plotada no eixo y
  - `hue="day"` – cor da agregação na coluna day

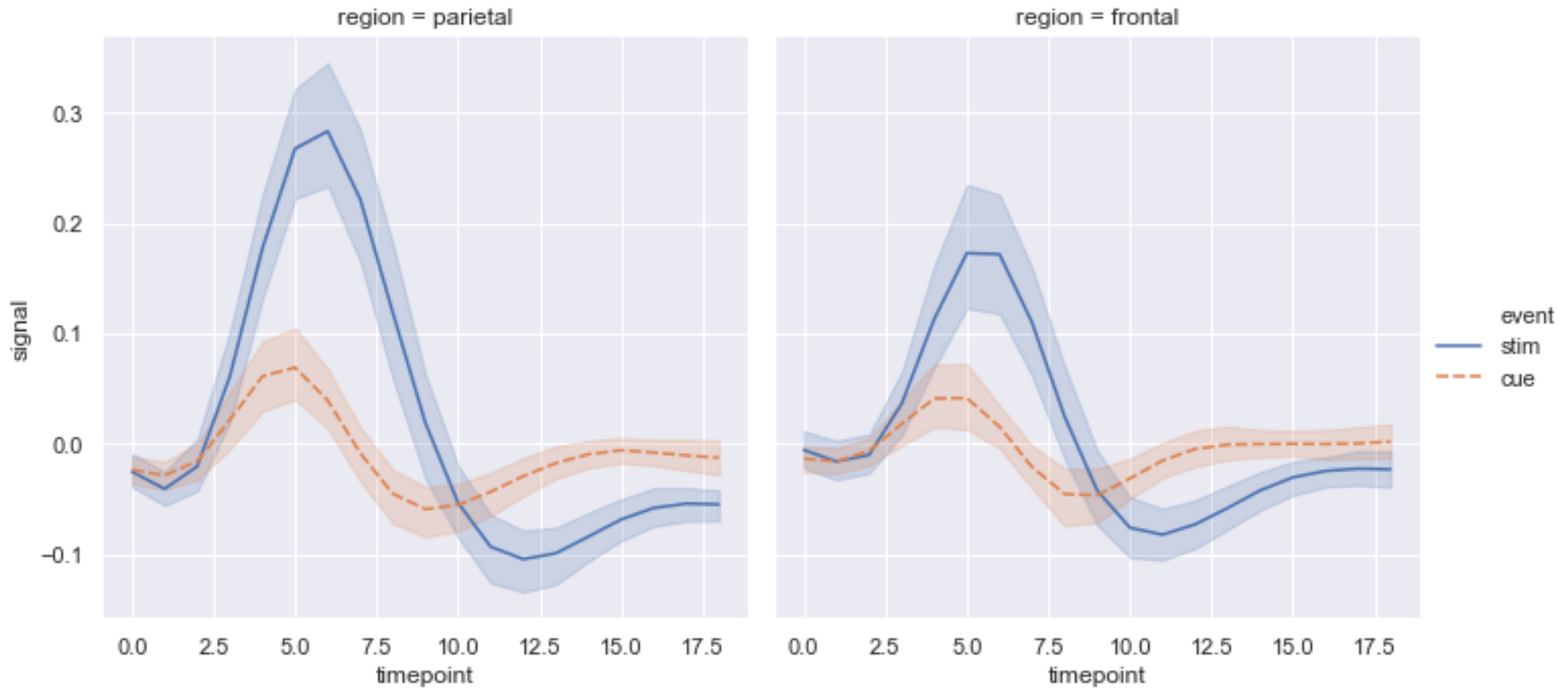
Dataset frmi

- Ressonância magnética do Cortex Cerebral

	subject	timepoint	event	region	signal
0	s13	18	stim	parietal	-0.017552
1	s5	14	stim	parietal	-0.080883
2	s12	18	stim	parietal	-0.081033
3	s11	18	stim	parietal	-0.046134
4	s10	18	stim	parietal	-0.037970

relplot de linha

```
fmri = sns.load_dataset("fmri")
sns.relplot(x="timepoint", y="signal", col="region",
            hue="event", style="event",
            kind="line", data=fmri);
```



relplot de linha

Using `kind="line"` offers the same flexibility for semantic mappings as `kind="scatter"`, but **`lineplot()`** transforms the data more before plotting.

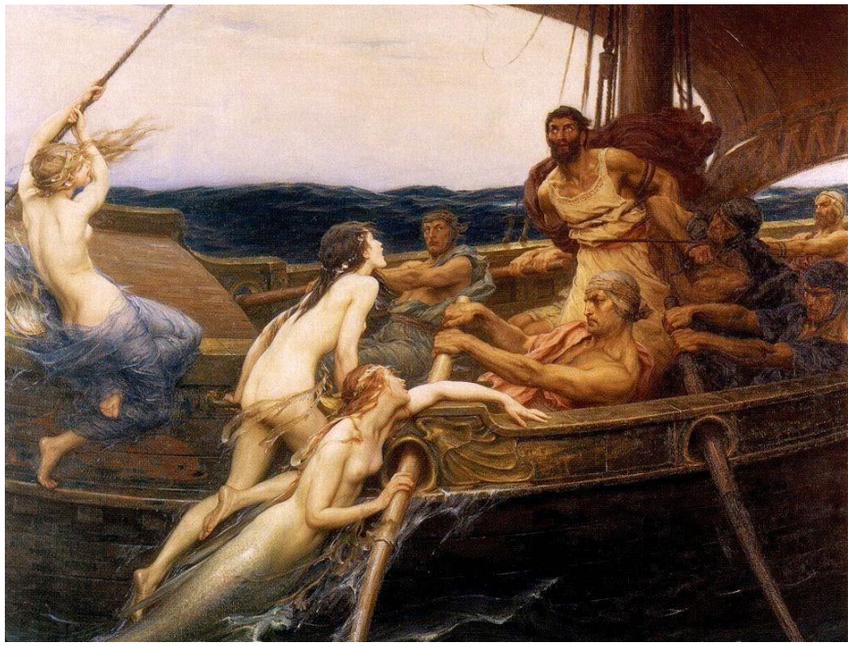
Observations are sorted by their `x` value, and repeated observations are aggregated.

By default, the resulting plot shows the mean and 95% CI for each unit

- As observações são ordenadas pelo argumento x
  - observações repetidas são agregadas
    - ✓ Calculando-se por padrão a média e intervalo de confiança de 95%
- O Seaborn vai além de plotar gráficos
  - Ele resolve problemas e aplica métodos estatísticos pra você
    - ✓ Intervalo de confiança, regressão linear, Kernel Density Estimation (kde), countplot

Aula Interativa

- É uma aula demonstrativa
  - A aula prática com tempo individual para resolver os exercícios
    - será depois
- Interaja respondendo as perguntas do Professor
  - Não tente resolver os exercícios no seu PC
    - Se preciso "amarre-se ao mastro do navio"
      - como Ulisses
- Atenção dividida
  - Resolva uma multiplicação e um Soma 1



Links Importantes

- Google Colab (Notebooks)
  - https://colab.research.google.com (Conta no Google/Gmail)
- Deepnote (Notebooks Multiusuário)
  - https://deepnote.com – Sem logar, só visualiza, não executa os códigos.
- Code Colab (Ambiente Python Multiusuário)
  - https://codecollab.io/@alexlopespereira/AulaEnap
- Slack (Link para entrar no workspace deste curso)
  - https://join.slack.com/t/enapespcd2021/shared_invite/zt-p8bygxfl-NxxzSWsSjZsIoyrDl70yXg
- Repositório da Disciplina no Github
  - https://github.com/alexlopespereira/enapespcd2021
- Dashboard de Notas
  - https://datastudio.google.com/s/htYxSs8wev4

Datas e Horários

- Horários
  - Manhã: 9h00 as 12h00
  - Tarde: 14h30 as 17h30
- Datas e Períodos
  - 8/11/2021 - Segunda-feira (Tarde)
  - 9/11/2021 - Terça-feira (Manhã e Tarde)
  - **16/11/2021 - Terça-feira (Manhã e Tarde)**
  - 22/11/2021 - Segunda-feira (Tarde)
  - 23/11/2021 - Terça-feira (Manhã e Tarde)
  - 29/11/2021 - Segunda-feira (Tarde)
  - 30/11/2021 - Terça-feira (Manhã e Tarde)
  - 6/12/2021 - Segunda-feira (Tarde)
  - 7/12/2021 - Terça-feira (Manhã e Tarde)

Material de Estudo

- Slides de Teoria
  - Com demonstrações em formato Gif
- Cadernos Colab
  - Teoria
  - Atividades
  - Exercícios
- Vídeos das Aulas

Avaliação

- Até 3 Exercícios por semana
  - Propostos na 3ª Feira (a tarde) e com deadline na 1ª aula seguinte
    - ✓ Em geral, uma 2ª Feira.
      - Exercícios atrasados valem 60% da nota.
- Convenção deste curso: Exercícios valem nota, Atividades não.
- Os exercícios das 3 primeiras semanas serão
  - Iguais para todos os alunos
- O exercício da penúltima semana será
  - De tema aberto (à conveniência dos alunos)
- Fórmula de Cálculo
  - Nota $= \sum_{i=1}^{K} \frac{\alpha \times N_i}{K}, \alpha = 0.6\ (se\ em\ atraso), K = numero\ de\ exercícios$

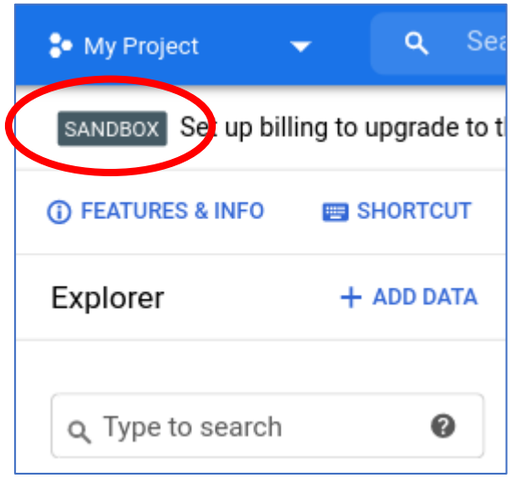
Configuração de Ambiente Python no Windows

- Instale o Python e o Pycharm no seu PC
  - Para a aula do dia 09 a tarde
- Opções de instalação do python
  - Anaconda (mais ferramentas)
    - ✓ Acesse https://docs.anaconda.com/anaconda/install/windows/ ou
    - ✓ O link direto pra versão 64 bits
      - https://repo.anaconda.com/archive/Anaconda3-2020.07-Windows-x86_64.exe
  - Somente o python (consome menos recursos do PC)
- Download e instalação do Pycharm (Versão **Community**)
  - https://www.jetbrains.com/pt-br/pycharm/download/#section=windows

BigQuery

- Data Warehouse
  - é um tipo de sistema de gerenciamento de dados projetado para ativar e fornecer suporte às atividades de Business Intelligence (BI).
- Solução Google de Armazenamento de Dados (Data Warehouse)
  - Gerenciado
  - Em escala de petabyte
  - Baixo Custo

BigQuery Sandbox



- Sandbox para teste sem cartão de crédito
  - https://www.youtube.com/watch?v=JLXLCv5nUCE
    - ✓ https://console.cloud.google.com/bigquery
  - Se não houver um projeto, crie um.
  - Query: 1TB/mês, Storage: 10TB/mês
- Sem billing account ativada
  - O ícone do sandbox aparece
    - ✓ Crie uma conta google ou encerre a billing account
      - Tutorial para encerrar billing account
        - https://cloud.google.com/billing/docs/how-to/manage-billing-account
- **Use o Sandbox!!**
  - Nem que pra isso seja necessário criar outra conta Google.

Base dos Dados – Projeto Open Source

- Acessar dados no BigQuery
  - Explicações em https://basedosdados.org
    - ✓ Acesso direto
      - https://console.cloud.google.com/bigquery?p=basedosdados&page=project
  - Projeto cuja missão é:
    - Organizar e mapear bases de dados brasileiras e internacionais.

Referências Bibliográficas

- KNAFLIC, C. N. (2018). Storytelling with data: a data visualization guide for business professionals.
- McKinney, W. (2018). Python for data analysis: Data wrangling with pandas, NumPy, and IPython.
- HURST, L. (2020). Hands on with Google Data Studio: a data citizen's survival guide. https://onlinelibrary.wiley.com/doi/book/10.1002/9781119616238.
- WEXLER, S., SHAFFER, J., & COTGREAVE, A. (2017). The big book of dashboards: visualizing your data using real-world business scenarios.
- https://www.storytellingwithdata.com/podcast
- https://seaborn.pydata.org/
- https://pandas.pydata.org/docs/
- https://d3js.org
- https://numpy.org/doc/stable/reference/index.html