MBA em Ciência de Dados

Técnicas Avançadas de Captura e Tratamento de Dados

Módulo III - Aquisição e Transformação de Dados

Coleta e Aquisição de Dados

Material Produzido por Moacir Antonelli Ponti

CeMEAI - ICMC/USP São Carlos

Conteúdo:

1. Bases de dados públicas
2. Coleta de dados

Referência complementar

DIEZ, David M.; BARR, Christopher D.; CETINKAYA-RUNDEL, Mine. **OpenIntro statistics**. 3.ed. OpenIntro, 2015. Capítulo 1.

Bases de dados públicas

Há **ferramentas de busca e sites** que indexam bases de dados que podem ser utilizados para provas de conceito ou análise iniciais em dados disponíveis publicamente

- Google Dataset Search : https://datasetsearch.research.google.com/ (https://datasetsearch.research.google.com/)
- OpenML : https://www.openml.org/search?type=data (https://www.openml.org/search?type=data)
- Kaggle : https://www.kaggle.com/datasets (https://www.kaggle.com/datasets)
- UCI : https://archive.ics.uci.edu/ml/index.php (https://archive.ics.uci.edu/ml/index.php)

No Brasil temos portais para **dados governamentais abertos**

- Federal: http://www.dados.gov.br/ (http://www.dados.gov.br/)
- Estado de São Paulo: http://catalogo.governoaberto.sp.gov.br/dataset (http://catalogo.governoaberto.sp.gov.br/dataset)
  - Dados educacionais: https://dados.educacao.sp.gov.br/ (https://dados.educacao.sp.gov.br/)
  - Outros indicadores: https://www.seade.gov.br/ (https://www.seade.gov.br/)
- Banco central do Brasil: https://dadosabertos.bcb.gov.br/ (https://dadosabertos.bcb.gov.br/)

Formatos comuns:

- Arquivos estruturados: XML, CSV, JSON, XLS, TXT
- Binários: PDF

Dados não estruturados:

- *feeds*
- mídias sociais
- sensores

Acesso via: arquivos ou APIs

(E)xtract, (T)ransform, (L)oad

É o processo comumente usado em datawarehouses, no qual Extract faz parte da aquisição de dados de diversas fontes.

Nesse módulo vamos também falar um pouco sobre o estágio de tranformação, comumente empregado, mas sem o contexto de ETL (que será visto posteriormente no curso)

Coleta de Dados

> *the general rule of thumb is: when in doubt, collect the data.*
>
> (a "regra de ouro" é: na dúvida, colete os dados)

Pontos importantes!

1. Conhecer as limitações éticas e legais

1. Projetar a coleta:
   - quais dados, seus tipos e como serão coletados?
   - por quanto tempo?
   - qual amostragem é necessária?

1. Considerar o tempo de implementação

Antes de começar

Limitações éticas e legais

Regulamentação de dados digitais

- Desde a GDPR européia implementada em 2018, a regulação de proteção de dados levou outros países a desenvolver mecanismos legais

  https://gdpr.eu/ (https://gdpr.eu/)

- No Brasil, iniciou como o conhecido **Marco Civil da Internet**
  - *Lei Geral de Proteção de Dados Pessoais (LGPD)*, lei 13.709 de 2018 http://www.planalto.gov.br/ccivil_03/_Ato2015-2018/2018/Lei/L13709.htm (http://www.planalto.gov.br/ccivil_03/_Ato2015-2018/2018/Lei/L13709.htm)
  - *Autoridade Nacional de Proteção de Dados (ANPD)*

Comitê de Ética

- A coleta de dados para pesquisa deve ser:
  1. cadastrada na plataforma Brasil http://plataformabrasil.saude.gov.br/ (http://plataformabrasil.saude.gov.br/)
  2. submetida a um comitê de ética em pesquisa

- Há diversos comitês de ética em pesquisa, como por exemplo:
  - UFSCar: http://www.propq.ufscar.br/etica/cep/humanos (http://www.propq.ufscar.br/etica/cep/humanos)
  - EACH/USP: http://www5.each.usp.br/apresentacao-cep/ (http://www5.each.usp.br/apresentacao-cep/)
  - Unifesp: https://cep.unifesp.br/ (https://cep.unifesp.br/)
  - Unicamp: https://www.prp.unicamp.br/pt-br/cep-comite-de-etica-em-pesquisa (https://www.prp.unicamp.br/pt-br/cep-comite-de-etica-em-pesquisa)

Planejando a coleta de dados

1. Como coletar?
    - implementar scripts de rastreamento (para *websites*)
    - montar Google Forms
    - implementar sistema de coleta de dados via:
        - aplicativo dedicado
        - jogos (sérios) ou via gamificação
        - redes sociais

1. O que coletar?
    - depende da pergunta que queremos responder
    - quando relacionado a um sistema, é útil listar todas as features do produto ou serviço
    - em negócios: listar todos os interessados (produto, engenharia, marketing, vendas)
        - relevante se os dados serão usados para KPI (indicador chave de desempenho)

1. Projetar a coleta:
    - especificar tipos, valores possíveis e como verificar integridade
    - entrevistar voluntários e envolvidos
    - realizar brainstroms e workshops
    - plataformas de coleta: rastreamento/tracking (transparente), questionário online (explicito, digital), questionário em papel (explícito, não digital)

1. Realizar a coleta
    - fazer estudo piloto para testar a coleta
    - acompanhar a coleta para evitar perda de dados
    - no caso de coleta em papel, recomenda-se dupla digitação com verificação de erros

Viés em bases de dados

> Good data > Big data

Um cuidado especial é evitar viés, em especial quando coletando ou analisando fontes de dados.

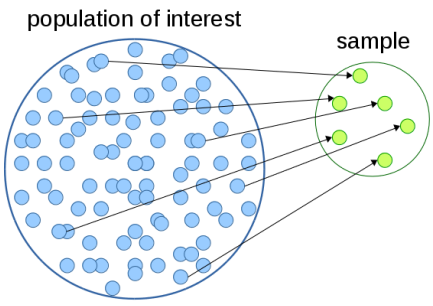
As fontes de viés mais comuns são:

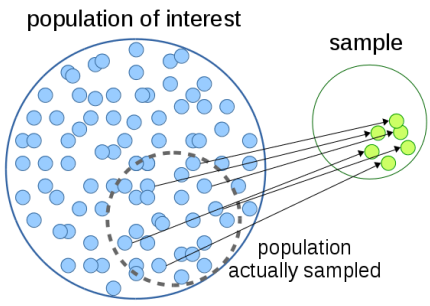
- Poucos dados
  - leva a evidências anetodais

- Viés de medida
  - uso de um instrumento errado
  - pergunta feita de forma errada
    - Lembrar as fontes de dados omissos!

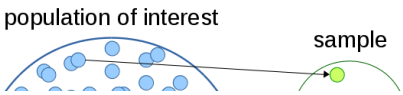
- Viés de amostragem
  - considerado frequentemente o mais perigoso
  - muitos métodos consideram premissas sobre a amostragem
  - frequentemente i.i.d. (independente e identicamente distribuída)

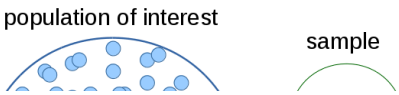
Amostragem de conveniência ou outro viés de seleção





(esq: amostragem representativa, dir: viés de seleção)





Resumo:

- Obter dados públicos é uma alternativa interessante para uma série de aplicações
  - é preciso lidar com diferentes formados e auditá-los
  - não estruturados são muito mais disponíveis, mas mais difícil de coletar e processar

- Coletar dados
  - exige **método** e planejamento
  - cuidado com questões legais e éticas
  - cuidado com amostragem

Conteúdo complementar

- TED series **can we trust the numbers?** https://www.npr.org/programs/ted-radio-hour/580617765/can-we-trust-the-numbers (https://www.npr.org/programs/ted-radio-hour/580617765/can-we-trust-the-numbers)

  Destaques:
  - *How can we tell the good statistics from the bad ones?* Mona Chalabi. https://www.npr.org/programs/ted-radio-hour/580617765/can-we-trust-the-numbers (https://www.npr.org/programs/ted-radio-hour/580617765/can-we-trust-the-numbers)
  - *Do Algorithms Perpetuate Human Bias?* Cathy O'Neil https://www.npr.org/2018/01/26/580617998/cathy-oneil-do-algorithms-perpetuate-human-bias (https://www.npr.org/2018/01/26/580617998/cathy-oneil-do-algorithms-perpetuate-human-bias)

- Williams, William H. *How bad can "good" data really be?*. The American Statistician 32.2 (1978): 61-65. https://www.researchgate.net/profile/Wm_Williams/publication/259529094_How_Bad_Can_Good_Data_Really_Be/links/0deec52c63bb2be889000000.pdf (https://www.researchgate.net/profile/Wm_Williams/publication/259529094_How_Bad_Can_Good_Data_Really_Be/links/0deec52c63bb2be889000000.pdf)