Gestão de Dados

Desafios estatísticos com dados de registros

Prof. Marcel de Toledo Vieira
Departamento de Estatística, ICE
UFJF

O que é o gerenciamento de dados?

- **Definição**: gerenciamento de dados é o processo de **organizar**, **armazenar**, **proteger** e **manter as informações** de uma organização de forma **eficiente** e **eficaz**.
  - envolve a **coleta**, **armazenamento**, **gerenciamento** e **distribuição de dados**, bem como a definição de políticas e procedimentos para garantir a **integridade**, **segurança** e **disponibilidade dos dados**.
- **Inclui**
  - o design de bancos de dados,
  - a implementação de sistemas de gerenciamento de bancos de dados,
  - a criação de políticas de segurança de dados,
  - a gestão do ciclo de vida dos dados e
  - a criação de estratégias de backup e recuperação de dados.

Para que?

- **Objetivos**
  - garantir que as informações da organização sejam precisas,
  - acessíveis,
  - seguras e
  - confiáveis para a tomada de decisões,
  - suporte a processos de negócios e
  - análises estatísticas.
- **O bom gerenciamento de dados**
  - reduz custos,
  - aumenta a eficiência e
  - aprimora a capacidade da organização de usar os dados para a tomada de decisões.

Um olhar da Estatística

- Dados administrativos estão se tornando cada vez mais importantes como **fonte de informação**, não apenas na educação mas também em outras áreas.
- Esses dados são geralmente um subproduto de alguma atividade operacional (**registros acadêmicos**, por exemplo) e tem vantagens sobre outras fontes.
- É necessária uma abordagem cautelosa sobre a sua coleta, manutenção e uso.
- Dados de registros trazem **desafios** específicos, e é importante que haja um debate sobre eles e esforços para melhorar sua **qualidade**.
- Pesquisadores de metodologia devem explorar esses problemas para ajudar a lidar com a crescente importância deste tipo de dados.

Registros Acadêmicos: *big data*

- Registros acadêmicos são gerados pelas próprias instituições e armazenados em **grandes bancos de dados**.
- Os conjuntos de dados de registros são geralmente enormes e impulsionaram o desenvolvimento tecnológico e o interesse em big data.
- *Big data* é um termo para conjuntos de dados tão grandes ou **complexos** que as aplicações tradicionais de processamento de dados são inadequadas.
- Os **desafios** incluem análise, captura, curadoria de dados, busca, compartilhamento, armazenamento, transferência, visualização, consulta, atualização e privacidade da informação.

BIG DATA

Big data

- *Big data* podem ser **estruturados** ou não estruturados.
- São gerados a uma velocidade cada vez maior.
- Podem ser analisados para a produção de **informações valiosas**.
- Esses dados podem vir de **diversas fontes**.
- Trabalhar com *big data* geralmente envolve o uso de técnicas avançadas de **estatística** e **computação**, o que leva a uma constante necessidade de capacitação dos envolvidos.

Tipos de B*ig data*

- **Social networks**
  - Facebook, Twitter, Blogs, Instagram, Picasa, YouTube, buscas no google etc), SMS, whatsapp, e-mails, ...
- **Registros administrativos** / dados de pesquisas
  - dados coletados por agências públicas, **registros acadêmicos**, registros médicos, e-commerce, dados bancários, ...
- **Dados gerados por máquinas**
  - dados de sensores de clima, poluição, etc, câmeras de trânsito e de vigilância, localização de telefones celulares, imagens de satélites, ...

Dados de registros administrativos

- A Organização para a Cooperação e Desenvolvimento Econômico (**OCDE**, 2016) define dados administrativos como tendo as seguintes características:
  - os dados foram originalmente coletados para um propósito não estatístico definido;
  - a **cobertura completa** da população-alvo é o objetivo;
  - o controle dos métodos pelos quais os dados administrativos são coletados e processados é de responsabilidade da instituição que coleta os dados.

Dados estatísticos *versus* dados de registros

- De acordo com Nordbotten (2010)
  - Dados estatísticos são coletados principalmente para fins estatísticos, como resumir para entender um fenômeno ou fazer **inferência**.
  - Dados de registros administrativos são coletados para algum **propósito administrativo**, para gerir uma organização, uma **universidade**, um **sistema educacional**, empresa, governo, escola, hospital.

Gestão de dados acadêmicos

- A **gestão acadêmica** pode exigir análises operacionais contínuas dos dados.
- Dados de registros idealmente consistem em dados de toda a população de **discentes**.
    - espera-se que tenham **alta qualidade** uma vez que o sucesso das atividades da instituição dependem disso.
    - podem ser utilizados para a produção de medidas resumo para o estudo de características descritivas da população.
    - precisam estar tão **atualizados** quanto for possível para representarem a instituição como ela é.

E na prática?

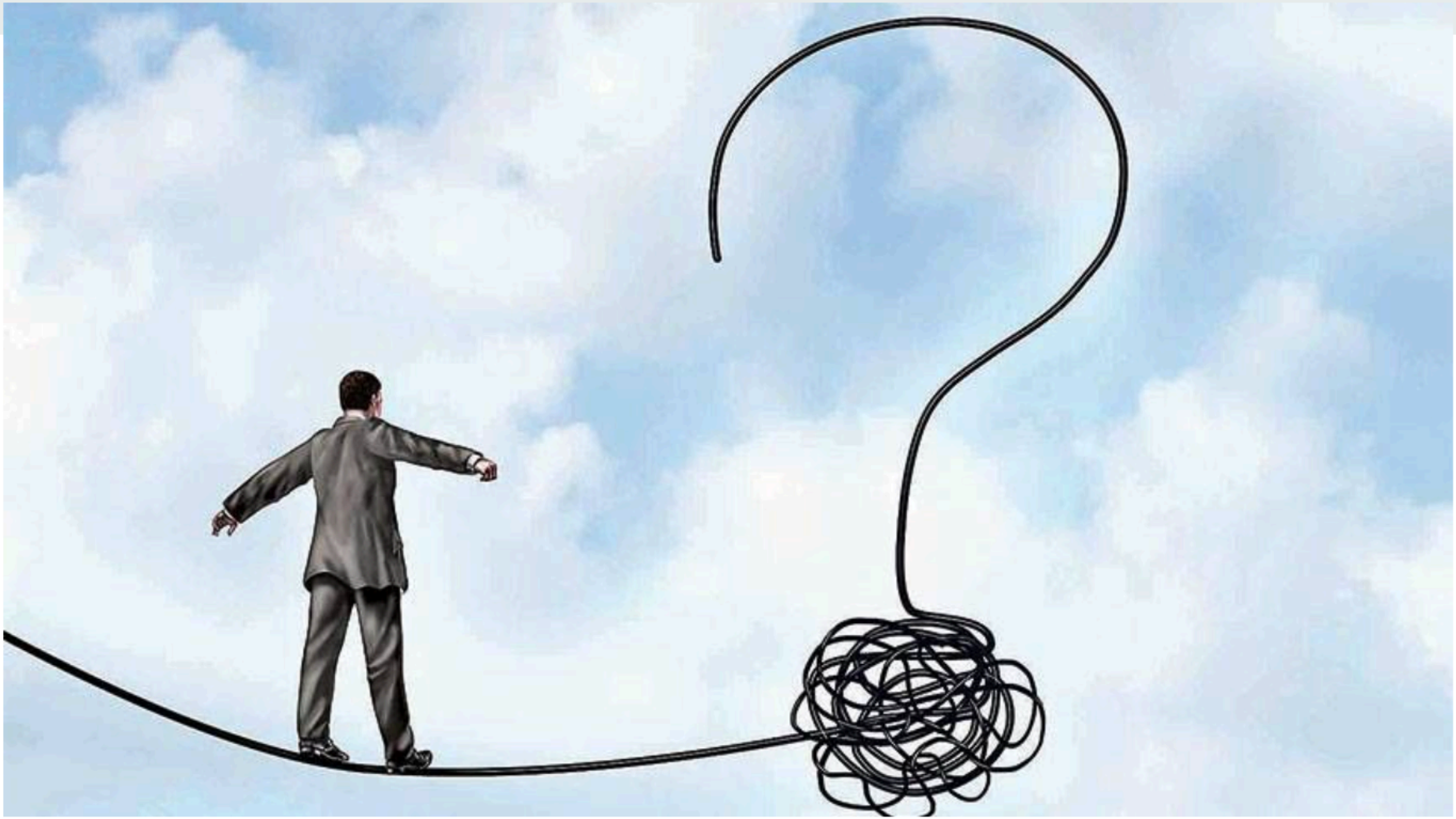
- Necessidades de **grande esforço institucional**, r**ecursos humanos**, **equipamentos**, **capacidade para o registro**, **limpeza**, **organização** e possivelmente vinculação a outros conjuntos de dados.
- **Desafio**: diferentes setores da instituição podem utilizar **diferentes** sistemas para registros, organização dos dados e bancos de dados.
- Na **prática**, **todos** os conjuntos de dados, especialmente aqueles que envolvem seres humanos, são susceptíveis a **problemas de qualidade**.
- **Dados de registros** não foram coletados para responder a questões estatísticas, mas suas definições e conceitos são uma boa aproximação.

Cuidados (I)

- Embora os dados de registros tenham méritos, devemos sempre abordar conjuntos de dados desse tipo com o mesmo **olhar crítico** com que abordamos qualquer outro conjunto de dados.
- **Caminho**: envolver estatísticos e analistas de dados na fase de coleta de dados de registros.
  - Eles podem ser capazes de **pensar adiante** e expandir o conjunto de dados coletados para que possam responder às perguntas futuras.
  - Necessidades de análises estatísticas futuras podem influenciar quais e como os **dados** de registros devem ser coletados e organizados.

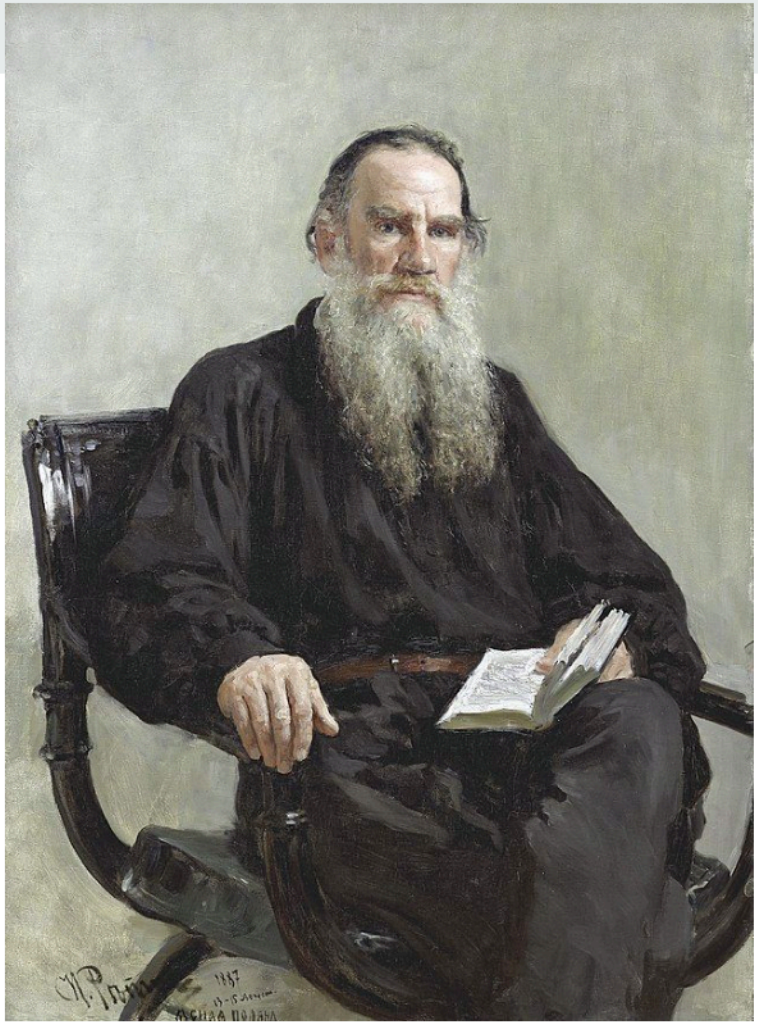
Cuidados (II)

- **Atenção**: dados de registros também estão sujeitos a **incertezas**.
- Os dados podem ser **'bons'** para um propósito, mas **'ruins'** para outro:
  - a **qualidade** não é uma propriedade do próprio conjunto de dados,
  - mas da interação entre o conjunto de dados e o uso ao qual ele é destinado.



Qualidade

- No entanto, a verdade é que os **dados de registros podem não ser nem completos nem livres de erros**.
- Quanto à "completude", a falta de informações pode se manifestar tanto na forma de **registros parciais**
  - registros nos quais alguns dos campos estão ausentes
  - quanto na forma de **registros inteiros faltantes**, de modo que o conjunto de dados não cubra toda a população de fato.
- Na verdade, **erros podem surgir de uma infinidade de maneiras**.



Parafraseando Leo **Tolstoy**:

- "***Um conjunto de dados perfeito é perfeito apenas de uma maneira; cada conjunto de dados imperfeito é imperfeito à sua própria maneira***".

Erros

- Nunca poderemos ter certeza de que todos os erros foram detectados em nosso conjunto de dados de registros.
- A maioria das **estruturas incomuns** em grandes conjuntos de dados surge de erros nos dados, em vez de qualquer coisa de interesse intrínseco.
- **Devemos desconfiar de qualquer conjunto de dados - grande ou pequeno - que pareça perfeito**.

Limpeza

- Os estatísticos sabem muito bem que é comum a maior parte do tempo ser dedicada à **limpeza de dados** antes de qualquer análise.
- **Desafio**: o processo de limpeza pode ser problemático quando o conjunto de dados é maciço, como nos registros acadêmicos de instituições médias e grandes.
- **Riscos de criar erros adicionais** durante processos automáticos e manuais de limpeza de dados.

Riscos

- Mecanismos de correção podem distorcer valores de dados perfeitamente bons, embora incomuns (atípicos).
- **Exemplo**: estratégia comum para lidar com valores ausentes é substituir pela média dos valores observados, o que nem sempre é o melhor.
- A familiaridade com o fato de que os dados muitas vezes não têm a melhor qualidade levou ao desenvolvimento de métodos e ferramentas estatísticas
  - métodos de detecção baseados em **verificações de integridade** e
  - em propriedades estatísticas.

Tipos de Erros

- Mesmo que os dados possam se afastar da qualidade perfeita de um número ilimitado de maneiras, é importante caracterizar o maior número possível de maneiras (Kim et al., 2003):
  - **erros de entrada de dados**,
  - **erros de atualização de dados**,
  - **erros de transmissão de dados** e
  - **bugs no sistema de processamento de dados**.
- Uma grande proporção de erros são de apenas alguns tipos, assim **um esforço relativamente pequeno levará a uma melhoria substancial**.


ERROS

Erros de entrada de dados

- Exemplos **clássicos** incluem
  - erros com datas,
  - dados sendo inseridos em colunas incorretas,
  - abreviações levando à confusões,
  - erros no uso de unidades de medida,
  - erros simples de ortografia,
  - etc.

Dimensões da Qualidade

- Principais **dimensões da qualidade**
    - precisão,
    - relevância,
    - pontualidade,
    - coerência,
    - completude,
    - acessibilidade e
    - segurança.

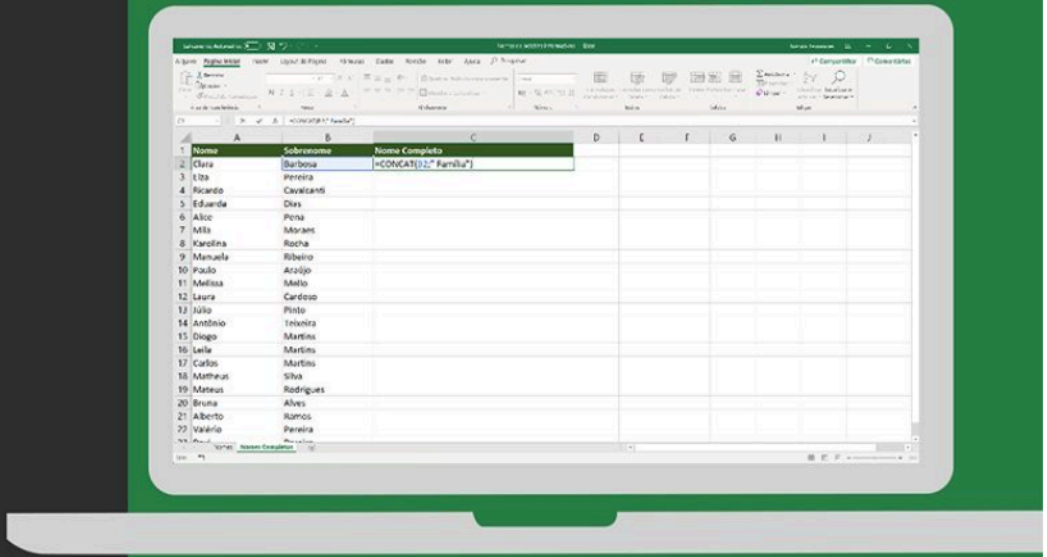
Dados de registros estão sempre completos?

- Muitos sistemas são dinâmicos e estão em constante mudança.
- Um banco de dados de todas as discentes de uma universidade hoje fornecerá, no máximo, apenas uma **imagem instantânea**.
- É certo que as discentes individuais terão mudado até o próximo ano, avançaram nos cursos, sem mencionar possíveis mudanças de nome devido a casamento e outros motivos, mudanças de endereço, etc.
- Essa **mudança populacional** apresenta desafios estatísticos interessantes e aponta para a fraqueza da afirmação de que os dados de registros representam "todos" os dados necessários.

Combinação de dados de diferentes fontes

- Combinar dados de **diferentes fontes** é cada vez mais importante.
- **Exemplo**: para fins estatísticos, como produzir resultados mais abrangentes.
- Ou informações são necessárias para uma organização de nível mais alto.
  - por exemplo, combinando estatísticas de várias instituições de ensino o **INEP** produz estatísticas para o Brasil como um todo.
- Desafios a serem enfrentados:
  - dados das diferentes instituições podem ter sido coletados usando métodos ou definições diferentes **->** combinações não são necessariamente simples.

Combinar dados



Motivos para combinar dados de diferentes fontes

(**a**) Diferentes fontes de dados e diferentes tipos de dados podem servir como complemento um ao outro, fornecendo diferentes tipos de informações.

(**b**) Triangulação e reconciliação entre fontes de dados são boas maneiras de verificarmos a qualidade dos dados de registros e corrigi-los se necessário.

- **Desafios**:
  - decidir quando combinar dois registros que não têm identificadores únicos e idênticos,
  - detecção de registros duplicados,
  - fusão de registros duplicados em uma única unidade.

Métodos para combinação de dados (Winkler, 2006)

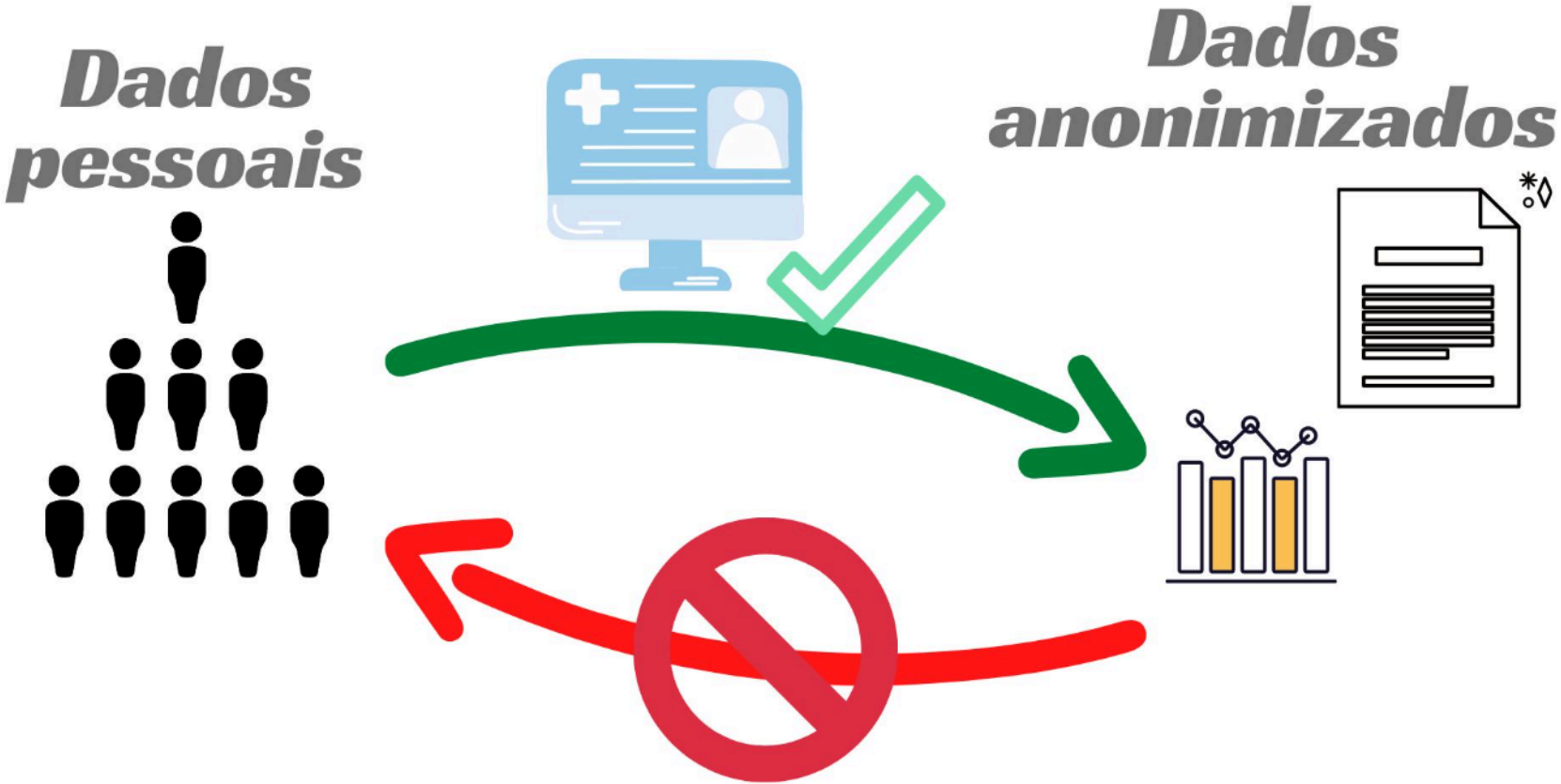
- **Correspondência manual**: não indicado para grandes bases de dados.
- Métodos **computacionais**: determinísticos e probabilísticos
  - **Métodos determinísticos**: verificam se dois registros concordam em um conjunto especificado de identificadores.
  - **Métodos probabilísticos**: relaxam o requisito de uma correspondência exata e, em vez disso, calculam uma medida de semelhança.
    - maximizam a probabilidade de uma correspondência.
- **Resultado**: correspondência, não correspondência ou possível correspondência.

Hiperdimensões da qualidade (Berka et al., 2012)

- A qualidade dos dados de registros pode ser avaliada em termos de **três** "hiperdimensões":
  - **documentação**: plausibilidade e aspectos legais,
  - **pré-processamento**: métodos formais para testar erros e inconsistências, e
  - comparação com uma fonte externa.
- Os resultados são três medidas, cada uma pontuada no intervalo 0 - 1.
- Uma média ponderada é tomada para produzir um **indicador de qualidade** geral.

Confidencialidade, privacidade e anonimização

- Um desafio comum com todos os dados que descrevem seres humanos é a necessidade de **preservar a confidencialidade e privacidade**.
- Essa questão é particularmente delicada com dados de registros, incluindo os de registros acadêmicos.
- **Alternativas**: existem ferramentas de anonimização e desidentificação
  - perturbação dos dados ou geração aleatória de dados com propriedades estatísticas iguais às dos dados brutos
- **Risco**: combinar um conjunto de dados com outros dados publicamente disponíveis para identificar um indivíduo e revelar algo sobre ele.


Dados pessoais
Dados anonimizados

Algumas considerações (I)

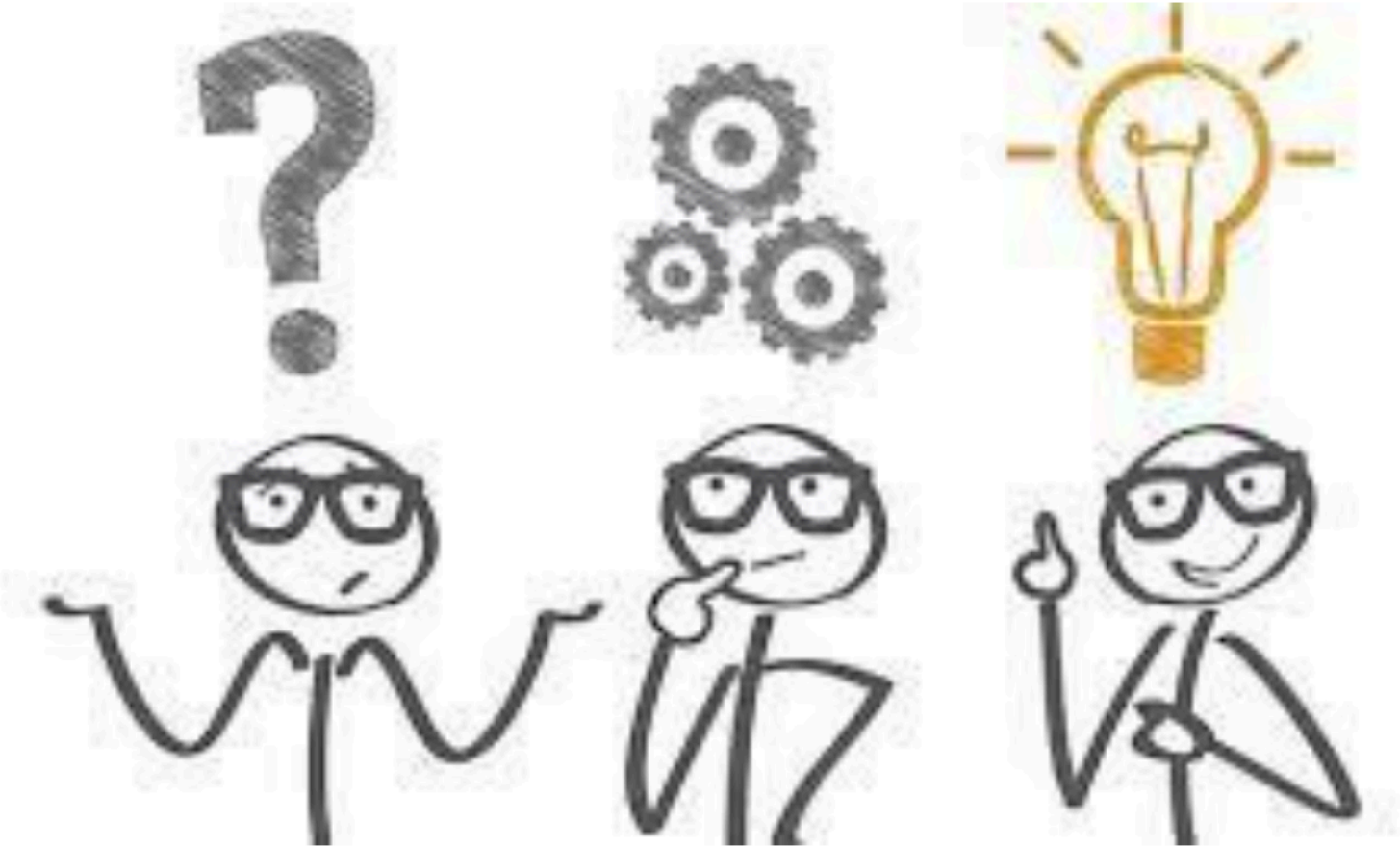
- Conjuntos de dados de registros são grandes e com boa cobertura populacional - embora possivelmente **vulneráveis** a outros problemas de qualidade.
- A **comunicação da incerteza**: precisamos saber comunicar as fontes de incerteza relacionadas aos dados de registros, uma vez que elas são muitas e variadas.
- **Educação estatística**: dados de registros estão se tornando tão importantes e tão amplamente utilizados, que se pode argumentar a favor de um ensino mais especializado de métodos específicos.

Algumas considerações (II)

- **Ambiente jurídico**: o aumento da conscientização sobre a moderna tecnologia de análise de dados estimulou culminou considerável pensamento legal e regulatório
  - consequência da privacidade e questões de confidencialidade.
- **No Brasil**: **Lei Geral de Proteção de Dados Pessoais** (LGPD), LEI Nº 13.709, DE 14 DE AGOSTO DE 2018, alterada pela Lei nº 13.853, de 2019.
  - Estabelece **regras para coleta, armazenamento, processamento e compartilhamento de dados pessoais, com o objetivo de proteger a privacidade.**

Considerações finais (I)

- Procurei estimular a discussão sobre a **necessidade de trabalhos metodológicos sobre dados de registros**.
- Esses dados estão sendo **usados cada vez mais amplamente** - em parte uma consequência da revolução do "big data".
- Os **problemas com este tipo de dados são diversos e heterogêneos**.



Considerações finais (II)

- Necessidade **urgente** de lidar com diferentes tipos de problemas de qualidade de dados, com o reconhecimento de que
  - não temos 'todos' os dados,
  - possíveis **incompatibilidades** entre a pergunta que queremos responder e as informações nos dados disponíveis,
  - a necessidade de **combinar dados** de múltiplas fontes bastante diferentes e
  - questões de **confidencialidade, privacidade e anonimização** que são muito desafiadoras.

marcel.vieira@ufjf.br