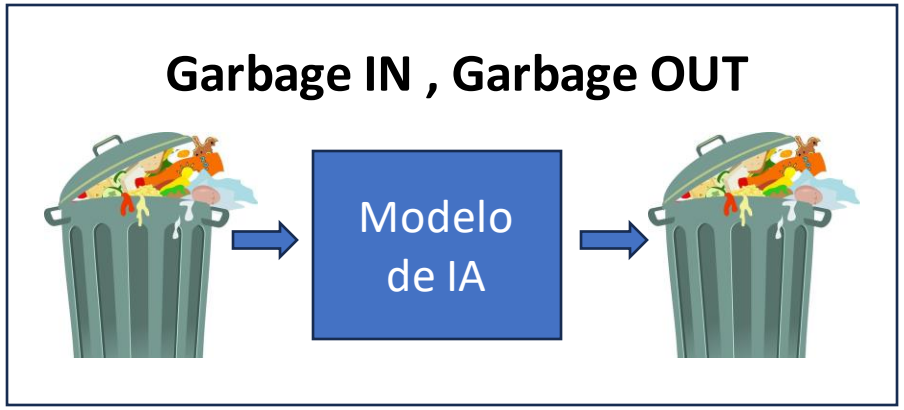
# Pré-processamento dos Dados

Profa Silvia Moraes

**Pré-processamento** é processo de preparação dos dados para o uso em algoritmos de aprendizado de máquina.

E é usado para:

- Melhorar a qualidade dos dados;
- Reduzir a complexidade computacional;
- Selecionar os atributos mais relevantes para aplicação;
- Adequar os dados aos formatos de entrada dos algoritmos;



**Técnicas para preparação de dados** podem ser agrupadas em:

- Eliminação manual de atributos
- Integração de dados;
- Amostragem de dados;
- Dados desbalanceados
- Limpeza de dados;
- Transformação de dados;
- Redução de dimensionabilidade.

# Eliminação manual de atributos

# Eliminação manual de atributos

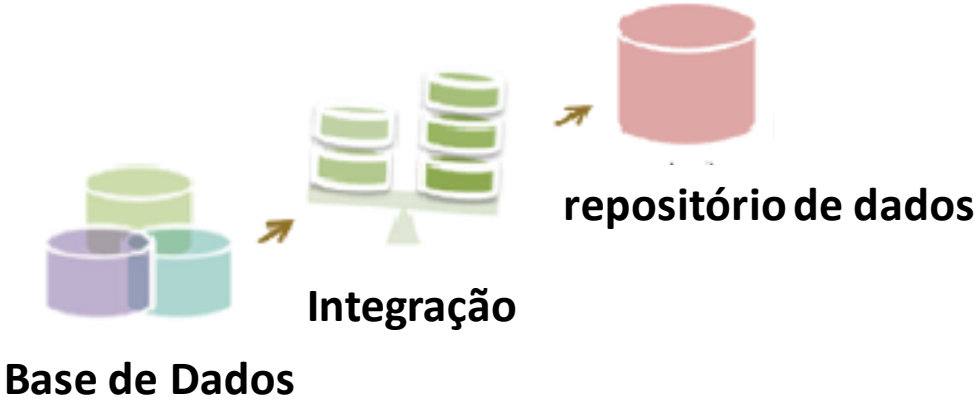
- Especialistas no domínio da aplicação podem analisar os dados e indicar os atributos mais relevantes;
- Existem **atributos irrelevantes** que são facilmente identificados e podem ser removidos já no inicio do pre-processamento.

| Id. | Nome | Idade | Sexo | Peso | Manchas | Temp. | # Int. | Est. | Diagnóstico |
|---|---|---|---|---|---|---|---|---|---|
| 4201 | João | 28 | M | 79 | Concentradas | 38,0 | 2 | SP | Doente |
| 3217 | Maria | 18 | F | 67 | Inexistentes | 39,5 | 4 | MG | Doente |
| 4039 | Luiz | 49 | M | 92 | Espalhadas | 38,0 | 2 | RS | Saudável |
| 1920 | José | 18 | M | 43 | Inexistentes | 38,5 | 8 | MG | Doente |
| 4340 | Cláudia | 21 | F | 52 | Uniformes | 37,6 | 1 | PE | Saudável |
| 2301 | Ana | 22 | F | ? | Inexistentes | 38,0 | 3 | RJ | Doente |
| 1322 | Marta | 19 | F | 87 | Espalhadas | 39,0 | 6 | AM | Doente |
| 3027 | Paulo | 34 | M | 67 | Uniformes | 38,4 | 2 | GO | Saudável |

# Integração de dados

# Integração de dados

- O processo de integração visa construir um repositório de dados: data warehouse;
- A integração geralmente exige tratamento dos dados:
  - Inconsistências
  - Ruídos
  - Exige transformação e correção



# Amostragem de dados

# Amostragem de dados

- Alguns algoritmos **tem dificuldades em lidar com muitos objetos**;
- **Quanto mais dados geralmente mais alta é a acurácia** (taxa de predições corretas). Porém isso **impacta no custo computacional**, deixando o tempo de processamento mais longo.
- Para reduzir o custo computacional, pode-se trabalhar com **amostras pequenas, mas representativas**.

# Amostragem de dados

- Abordagens que podem ser usadas para amostragem:
  - **Aleatória simples**: busca exemplos dos dados originais de forma aleatória;
  - **Estratificada**: busca exemplos de cada classe de forma proporcional a quantidade original ou mantendo o mesmo numero de amostras;
  - **Progressivas**: comeca com uma amostra pequena e vai aumentando ao longo do desenvolvimento;

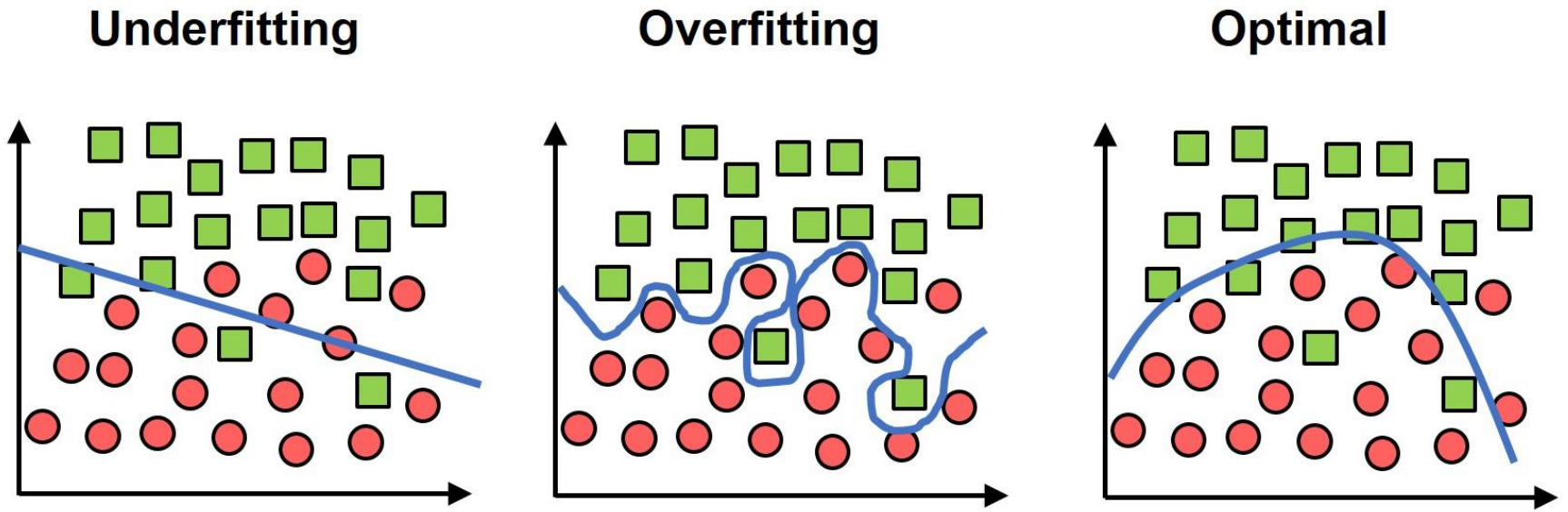
# Dados desbalanceados

# Dados desbalanceados

- Vários algoritmos de machine learning tem seu desempenho prejudicado por causa de dados desbalanceados;
- Os algoritmos tendem a favorecer a classe majoritária, com mais amostras;
- Técnicas para amenizar o desbalanceamento:
  - Redefinir o tamanho do conjunto de dados;
  - Uso de modelos binários para reconhecer as classes;

# Dados desbalanceados

- Podem provocar:
  - **Overfitting**(sobreajuste ou superajuste) e **Underfitting**(sub-ajuste) em Machine Learning são conceitos que se referem ao ajuste do modelo.




https://datahacker.rs/018-pytorch-popular-techniques-to-prevent-the-overfitting-in-a-neural-networks/

# Limpeza dos Dados

# Limpeza de dados

- Muitos dados do mundo real são potencialmente incorretos (falha no instrumento de leitura, erro humano ou de máquina, não obrigatoriedade, erro de transmissão).
- Os dados podem ser:
  - **Incompletos**: falta de valores de atributos, falta de certos atributos de interesse ou contendo apenas dados agregados. Por exemplo, Ocupação = "" (dados faltantes) .
  - **Ruidosos:** contendo ruído, erros ou outliers. Por exemplo, Salário = "- 10" (um erro)
  - **Inconsistentes:** contendo discrepâncias em códigos ou nomes. Exemplos: Idade = "42" e Aniversário = "03/07/2010"; ora a é classificação "1, 2, 3" ora é "A, B, C"; “01 de janeiro” como o aniversário de todos? CEP de todos 90000-000?
  - **Redundantes:** pode ter objetos e atributos redundantes

# Limpeza de dados

- O que fazer com **dados incompletos** ?

| Id. | Nome | Idade | Sexo | Peso | Manchas | Temp. | # Int. | Est. | Diagnóstico |
|---|---|---|---|---|---|---|---|---|---|
| 4201 | João | 28 | M | 79 | Concentradas | 38,0 | 2 | SP | Doente |
| 3217 | Maria | 18 | F | 67 | Inexistentes | 39,5 | 4 | MG | Doente |
| 4039 | Luiz | 49 | M | 92 | Espalhadas | 38,0 | 2 | RS | Saudável |
| 1920 | José | 18 | M | 43 | Inexistentes | 38,5 | 8 | MG | Doente |
| 4340 | Cláudia | 21 | F | 52 | Uniformes | 37,6 | 1 | PE | Saudável |
| 2301 | Ana | 22 | F | ? | Inexistentes | 38,0 | 3 | RJ | Doente |
| 1322 | Marta | 19 | F | 87 | Espalhadas | 39,0 | 6 | AM | Doente |
| 3027 | Paulo | 34 | M | 67 | Uniformes | 38,4 | 2 | GO | Saudável |

# Limpeza de dados

- O que fazer com **dados incompletos** ?
  - **Eliminar os objetos com dados faltantes**:
    - Pode ser usada quando o dado faltante é a classe;
    - Não é indicada quando
      - poucos atributos do objeto possuem valores ausentes;
      - o conjunto de dados é pequeno;
      - o dado faltante é diferente para cada objeto;
  - **Definir e preencher manualmente**: tedioso e inviável para muitos dados.

# Limpeza de dados

- O que fazer com **dados incompletos** ?
  - ...
  - **Preencher automaticamente** (uso de algum método ou heurística):
    - uma constante global: por exemplo, "desconhecido"
    - média : a média ou mediana do atributo para todas as amostras pertencentes à mesma classe ou moda, em caso de valor simbólico (uma boa opção)
    - valor mais provável: baseado em inferência (uso de uma fórmula bayesiana ou árvore de decisão)

# Limpeza de dados

- O que fazer com **dados ruidosos** ?

| Id. | Nome | Idade | Sexo | Peso | Manchas | Temp. | # Int. | Est. | Diagnóstico |
|---|---|---|---|---|---|---|---|---|---|
| 4201 | João | 28 | M | 79 | Concentradas | 38,0 | 2 | SP | Doente |
| 3217 | Maria | 18 | F | 67 | Inexistentes | 39,5 | 4 | MG | Doente |
| 4039 | Luiz | 49 | M | 92 | Espalhadas | 38,0 | 2 | RS | Saudável |
| 1920 | José | 18 | M | 43 | Inexistentes | 38,5 | 8 | MG | Doente |
| 4340 | Cláudia | 21 | F | 52 | Uniformes | 37,6 | 1 | PE | Saudável |
| 2301 | Ana | 22 | F | 3000 | Inexistentes | 38,0 | 3 | RJ | Doente |
| 1322 | Marta | 19 | F | 87 | Espalhadas | 39,0 | 6 | AM | Doente |
| 3027 | Paulo | 34 | M | 67 | Uniformes | 38,4 | 2 | GO | Saudável |

# Limpeza de dados

- O que fazer com **dados ruidosos** ?

- **Encestamento**
  1. Classificar os dados e organizá-los em cestas ou faixas (de frequência igual)
  2. Suavizar o ruído, substituindo os valores pela média ou mediana dos valores pertencentes à mesma faixa de valor.
- **Agrupamento**: detectar e remover outliers (atributos que não formarem grupos)
- **Regressão**: Ajustando os dados por meio de funções de regressão e por classificação, no caso de dados simbólicos.
- **Distância**: técnicas baseadas em distância verificam a que classe pertencem objetos mais próximos a x. Se x for de outra classe, ele pode ser um ruído. Borderlines devem ser eliminados.

# Limpeza de dados

- O que fazer com **dados inconsistentes** ?

| Id. | Nome | Idade | Sexo | Peso | Manchas | Temp. | # Int. | Est. | Diagnóstico |
|---|---|---|---|---|---|---|---|---|---|
| 4201 | João | 28 | M | 79 | Concentradas | 38,0 | 2 | SP | Doente |
| 3217 | Maria | 18 | F | 67 | Inexistentes | 39,5 | 4 | MG | Doente |
| 4039 | Luiz | 49 | M | 92 | Espalhadas | 38,0 | 2 | RS | Saudável |
| 1920 | José | 18 | M | 43 | Inexistentes | 38,5 | 8 | MG | Doente |
| 4340 | Cláudia | 21 | F | 52 | Uniformes | 37,6 | 1 | PE | Saudável |
| 2301 | Ana | 18 | F | 67 | Inexistentes | 39,5 | 4 | RJ | Saudável |
| 1322 | Marta | 19 | F | 87 | Espalhadas | 39,0 | 6 | AM | Doente |
| 3027 | Paulo | 34 | M | 67 | Uniformes | 38,4 | 2 | GO | Saudável |

?

?

# Limpeza de dados

- O que fazer com **dados inconsistentes** ?

- Como geralmente são dados gerados por processos de integração e que violam regras de relações conhecidas, pode-se utilizar algoritmos para analisar os dados e identificar as inconsistências.
  - escalas diferentes para uma mesma medida (m, cm)
  - codificação diferente para representar um atributo relacionado a tamanho (pequeno e grande; médio e enorme).

# Limpeza de dados

- O que fazer com **dados inconsistentes** ?
  - Podem ser identificados pelo cálculo de correlação (mede o quanto duas variáveis tendem a mudar juntas) e análise de covariância (mede a relação linear entre duas variáveis).

$$\text{Covariancia(x,y)} = 1/(n-1)\sum_{i=1}^{n}\left((x_i - \bar{x}_i)(y_i - \bar{y}_i)\right)$$

$$\text{Correlação } \rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

  - Quando as inconsistências não puderem ser corrigidas, esses dados devem ser removidos.

# Limpeza de dados

- O que fazer com **dados redundantes** ?

| Id. | Nome | Idade | Sexo | Peso | Manchas | Temp. | # Int. | Est. | Diagnóstico |
|---|---|---|---|---|---|---|---|---|---|
| 4201 | João | 28 | M | 79 | Concentradas | 38,0 | 2 | SP | Doente |
| 3217 | Maria | 18 | F | 67 | Inexistentes | 39,5 | 4 | MG | Doente |
| 4039 | Luiz | 49 | M | 92 | Espalhadas | 38,0 | 2 | RS | Saudável |
| 1920 | José | 18 | M | 43 | Inexistentes | 38,5 | 8 | MG | Doente |
| 4340 | Cláudia | 21 | F | 52 | Uniformes | 37,6 | 1 | PE | Saudável |
| 2301 | Ana | 18 | F | 67 | Inexistentes | 39,5 | 4 | MG | Doente |
| 1322 | Marta | 19 | F | 87 | Espalhadas | 39,0 | 6 | AM | Doente |
| 3027 | Paulo | 34 | M | 67 | Uniformes | 38,4 | 2 | GO | Saudável |

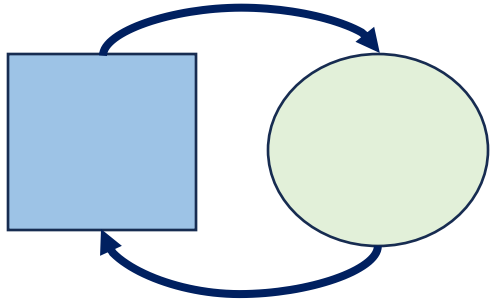
?

?

# Limpeza de dados

- O que fazer com **dados redundantes** ?
  - **Eliminar dados redundantes**: tuplas cujos atributos possuem os mesmos valores (ou muito próximos).
  - **Eliminar atributos redundantes** (atributos que podem ser deduzidos a partir de outros). Ex: idade e data de nascimento; quantidade de vendas, valor por venda e venda total

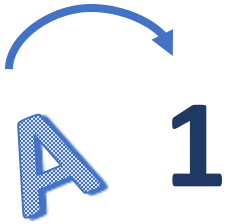
# Transformação de dados



# Transformação de dados

- Algumas técnicas em aprendizado de máquina só trabalham com um tipo de dado: apenas numérico ou apenas simbólico.
- As **transformações** pode ser:
  - Conversão Simbólico-Numérico
  - Conversão Numérico-Simbólico
  - Normalização
  - Simplificação

# Transformação de dados



- **Conversão Simbólico-Numérico**
  - Algoritmos como Redes Neurais só trabalham com valores numéricos.

| Valor ordinal | Valor inteiro |
|---|---|
| Primeiro | 1 |
| Segundo | 2 |
| Terceiro | 3 |
| Quarto | 4 |
| Quinto | 5 |

| Valor ordinal | Valor inteiro |
|---|---|
| Pequeno | 1 |
| Medio | 2 |
| Grande | 3 |

| Valor ordinal | Valor inteiro |
|---|---|
| Positivo | +1 |
| Neutro | 0 |
| Negativo | -1 |

| Estado Civil | Código |
|---|---|
| Solteiro | 1 |
| Casado | 2 |
| Divorciado | 3 |
| Desquitado | 4 |
| Viúvo | 5 |

| Estado | Sigla | Código |
|---|---|---|
| Rondônia | RO | 11 |
| Acre | AC | 12 |
| Amazona | AM | 13 |
| Roraima | RR | 14 |
| Para | PA | 15 |
| ... | | |

# Transformação de dados


- **Conversão Numérico-Simbólico**

  - Alguns algoritmos trabalham apenas com dados qualitativos.
  - Se o valor discreto ou binário, esta transformação e' simples.

  - Algumas estratégias de transformação:
    - **Larguras iguais:** divide o intervalo original de valores em subintervalos com mesma largura. (outliers podem prejudicar essa estratégia)
    - **Frequências iguais**: divide o intervalo original por frequência (pode gerar subintervalos de tamanhos bem diferentes).
    - **Uso de algum algoritmo de agrupamento**
    - **Inspeção Visual**

# Transformação de dados

- **Normalização**
  - Recomendada quando os limites de valores de atributos distintos são muito diferentes;
  - Evita que um atributo predomine sobre outro;
  - A normalização pode ser por **amplitude** ou **distribuição**:
    - **Distribuição:** muda a escala de valores de um atributo. Ex: ordena os valores e substitui seus valores pela sua posição no ranking. (Valores: 19,8,7,2,7; substitui por 4,3,2,1,2)
    - Se todos os valores forem distintos, a distribuição é uniforme

# Transformação de dados

- **Normalização**
  - **Amplitude:** pode ser por reescala ou padronização (padronização lida melhor com outliers).
    - **Reescala**: define uma nova escala, com limites mínimo (min) e máximo(max) novos para todos os atributos. Menor e maior são os limites atuais.

$$valor_{novo} = min + \frac{valor_{atual} - menor}{maior - menor}(max - min)$$

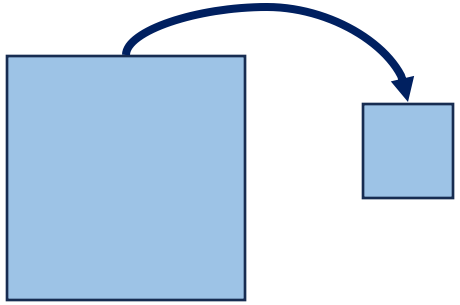
    - **Padronização:** define um valor central e um valor de espalhamento comuns a todos os atributos (σ= covariancia e μ=media)

$$valor_{novo} = \frac{valor_{atual} - \mu}{\sigma}$$

# Transformação de dados

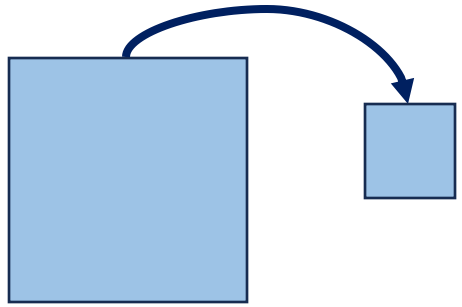
- **Simplificação**:
  - transformação para um valor mais facilmente manipulável.
  - Ex: idade ao invés de data de nascimento.

# Redução de dimensionabilidade

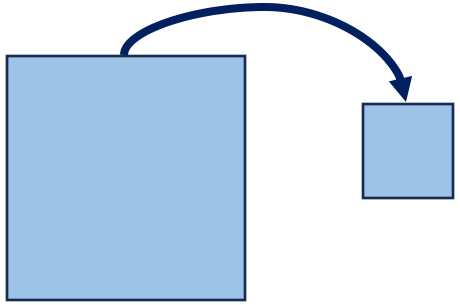


# Redução de dimensionabilidade

- Muitos problemas possuem um número elevado de atributos, especialmente textos e imagens.
- As técnicas que visam reduzir dimensionabilidade seguem as abordagens:
  - agregação
  - seleção de atributos

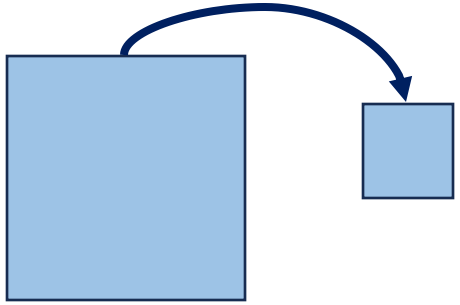


# Redução de dimensionabilidade



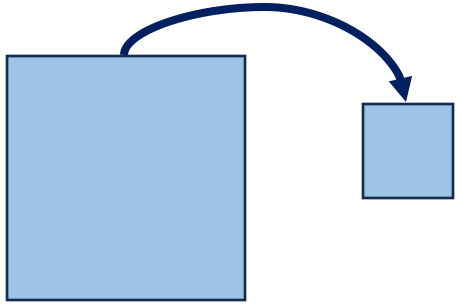
- **Agregação**: combina os atributos originais por meio de funções lineares ou não lineares.
  - **Análise de Componentes Principais** (PCA - Principal Component Analysis): técnica bem conhecida que correlaciona estatisticamente os exemplos, reduzindo a dimensionalidade do conjunto de dados original pela eliminação de redundâncias.
  - Obs: Essa técnica leva a perda dos valores originais. Em várias aplicações (áreas de biologia, finanças, medicina, etc), os valores originais são importantes para a interpretação dos resultados. Por isso, técnicas de seleção de atributos em determinadas areas são mais usadas.

# Redução de dimensionabilidade



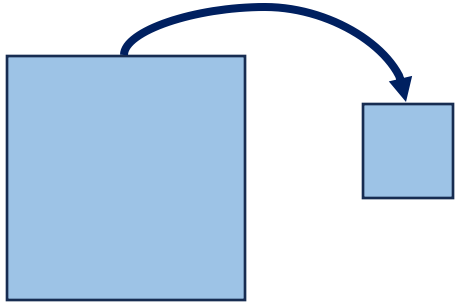
- A **seleção de atributos** busca um subconjunto ótimo de atributos para o problema.
- Ela permite:
  - identificar atributos importantes;
  - melhorar o desempenho dos algoritmos de aprendizado;
  - reduzir exigência de memória e processamento;
  - eliminar atributos irrelevantes e ruídos;
  - simplificar o modelo gerado e, consequentemente, sua compreensão;
  - facilitar a visualização dos dados;

# Redução de dimensionabilidade



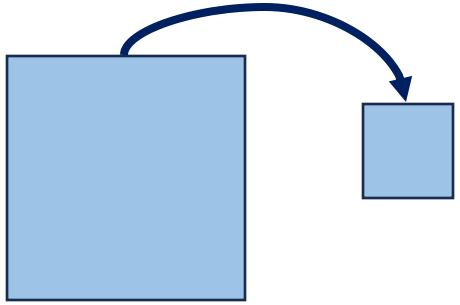
- A **seleção de atributos** não é uma tarefa trivial, pois pode existir:
  - número muito grande de exemplos;
  - número muito grande de atributos;
  - relações complexas entre atributos, que dificultam a descoberta de relações entre eles.

# Redução de dimensionabilidade



- Abordagens usadas para **seleção de atributos**:
  - **Embutida:** o próprio algoritmo define os atributos. Ex: arvore de decisão.
  - **Baseada em filtro**: podem usar a correlação entre os dados como critério de seleção.
  - **Baseada em wrapper**: nessa abordagem o algoritmo de classificação é executado para cada subconjunto e a avaliação geralmente é feita em termos da acurácia preditiva retornada pelo algoritmo.O subconjunto que trouxer o melhor desempenho de aprendizado é definido como o melhor subconjunto de atributos.
    - Estratégias: Backward Generation(inicia com todos e remove um por vez); Forward Generation(vai acrescentando um por vez); Estocástica; ...

# Redução de dimensionabilidade



- Existem várias técnicas que visam selecionar atributos, as mais simples são baseadas em ordenação.
- Ordena de acordo com algum critério (exemplo frequência) e seleciona por:
  - **Ranking:** escolhe os n primeiros
  - **Relevância:** escolhe todos os atributos cujo valor está acima de um limiar n.