Universidade do Minho
Departamento de Informática

APRENDIZAGEM E DECISÃO INTELIGENTES

LEI/MiEI @ 2022/2023, 2º sem [ADI^3]

ISLab
Synthetic Intelligence Lab

Agenda

- Porquê Preparação de Dados?
- Tarefas
  - Discretização
  - Limpeza
  - Integração
  - Transformação
  - Redução
- Tipos de dados
  - Qualitativos
  - Quantitativos

ISLab
Synthetic Intelligence Lab

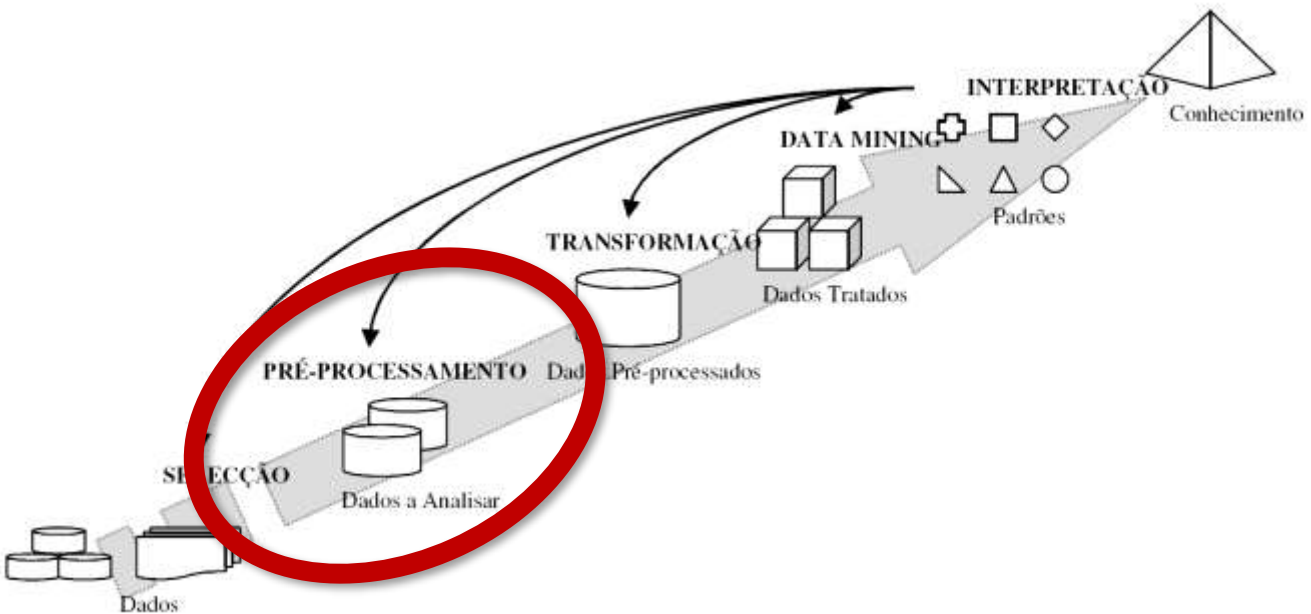
Porquê preparar os dados?

- **Porque SIM!**
- O principal objetivo da **preparação dos dados consiste em transformar** os *data sets* por forma a que a **informação** neles contida esteja **adequadamente exposta à ferramenta** de análise de dados (AD);
- A preparação dos dados "também prepara o preparador" por forma selecionar os modelos de AD mais adequados;
- Os dados têm de ser formatados para se adequarem a uma determinada ferramenta de AD;
- Os dados recolhidos do "mundo real":
  - são incompletos;
  - contêm lixo;
  - podem conter inconsistências.

ISLab
Synthetic Intelligence Lab

Preparação de dados

- Preparação dos Dados (Pré-processamento).
  Como?
  - Discretização;
    (classes etárias)
  - Limpeza;
    (nº BI)
  - Integração;
    (fontes)
  - Transformação;
    (diários/mensais)
  - Redução de dados.
    (moradas/regiões)



Data Mining – Descoberta de Conhecimento em Bases de Dados
Manuel Filipe Santos, Carla Azevedo

ISLab
Synthetic Intelligence Lab

Porquê preparar os dados?

- Os dados recolhidos do "mundo real":
  - são incompletos:
    - falta de valores em alguns atributos;
    - falta de alguns atributos;
    - dados agregados ou generalizados;
    - Código postal: 4710-... Braga;
    - N° de filhos: "";
  - contêm lixo;
  - podem conter inconsistências.

ISLab
Synthetic Intelligence Lab

Porquê preparar os dados?

- Os dados recolhidos do “mundo real”:
  - são incompletos;
  - contêm lixo:
    - identificam valores impossíveis;
    - Salário: -1.000EUR;
    - Idade: 321;
    - Data: 31/novembro/2017;
    - País: Catalunha;
  - podem conter inconsistências.

ISLab
Synthetic Intelligence Lab

Porquê preparar os dados?

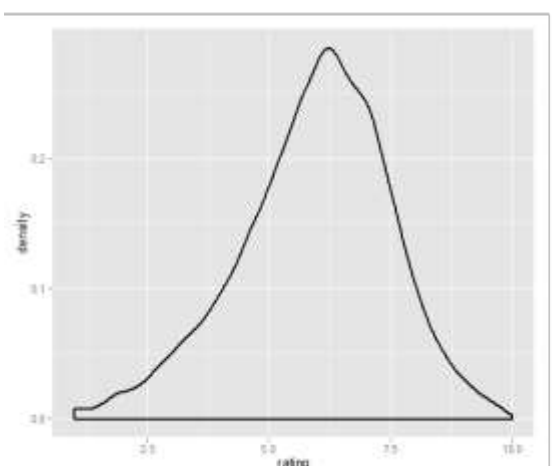
- Os dados recolhidos do “mundo real”:
  - são incompletos;
  - contêm lixo;
  - podem conter inconsistências:
    - encontram-se discrepâncias entre valores ou nomes;
    - Idade = 35; Data de nascimento = 31/maio/1969;
    - Sexo: “M/F”; “0/1”; “Masculino/Feminino/Desconhecido”;
    - diferenças entre valores de registos duplicados.

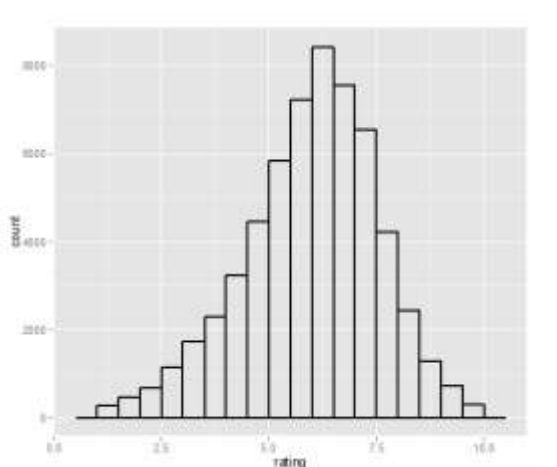
https://www.metropoles.com/brasil/rape-sinalizacao-errada-em-rua-no-interior-de-sp-vira-piada

ISLab
Synthetic Intelligence Lab

Tarefas na preparação de dados

- Discretização/Enumeração:
  - Redução de dados com importante aplicação a dados numéricos;
- Limpeza;
- Integração;
- Transformação;
- Redução.





ISLab
Synthetic Intelligence Lab

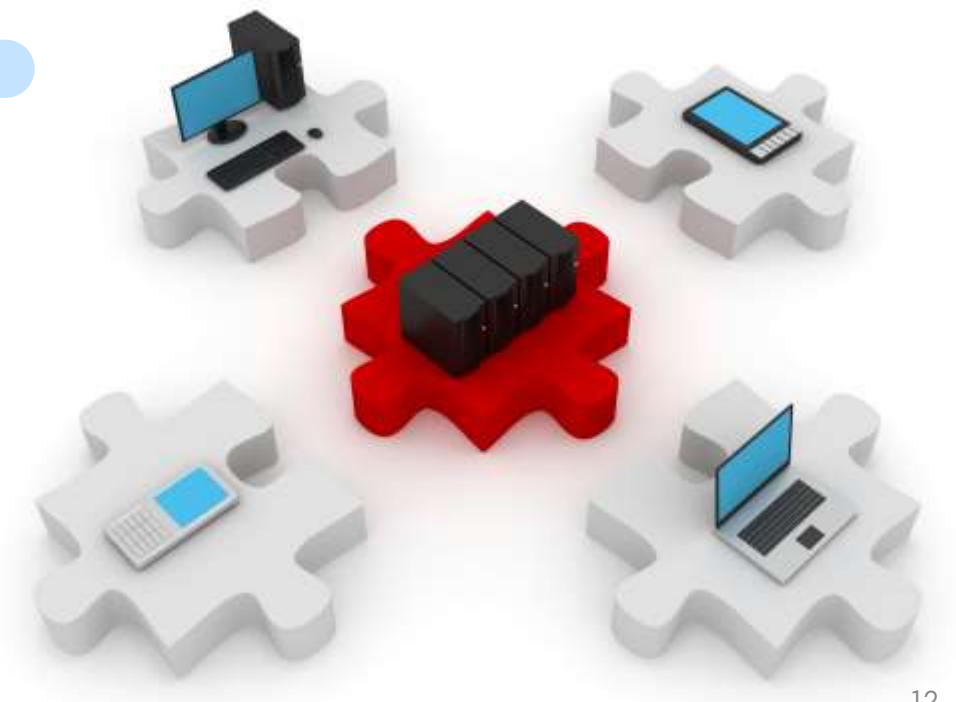
Tarefas na preparação de dados

- Discretização/Enumeração;
- Limpeza:
  - Preenchimento de valores de atributos;
  - Remoção de lixo dos dados;
  - Remoção de valores impossíveis;
  - Resolução de inconsistências;
- Integração;
- Transformação;
- Redução.

ISLab
Synthetic Intelligence Lab

Tarefas na preparação de dados

- Discretização/Enumeração;
- Limpeza;
- Integração:
  - Múltiplas fontes de dados (BD's, ficheiros, papel, web, etc.);
- Transformação;
- Redução.



ISLab
Synthetic Intelligence Lab

Tarefas na preparação de dados

- Discretização/Enumeração;
- Limpeza;
- Integração;
- Transformação:
  - Normalização e agregação de dados;
- Redução.

ISLab
Synthetic Intelligence Lab

Tarefas na preparação de dados

- Discretização/Enumeração;
- Limpeza;
- Integração;
- Transformação;
- Redução:
  - Obtenção de representações de dados menos volumosas, mas com capacidade para produzir idênticos resultados analíticos;
  - Redução de dimensões;
  - Compressão de dados.

Tipos de dados

- Os tipos dos dados diferem na sua natureza e na quantidade de informação que proporcionam:
- **Qualitativos ou Quantitativos**.

ISLab
Synthetic Intelligence Lab

Tipos de dados

- Nominais:
  - Atribui nomes únicos a objetos:
    - Não existe outra informação que se possa deduzir;
    - Nomes de pessoas;
    - Códigos de identificação;
- Categorias;
- Ordinais;
- Intervalos;
- Rácios.

ISLab
Synthetic Intelligence Lab

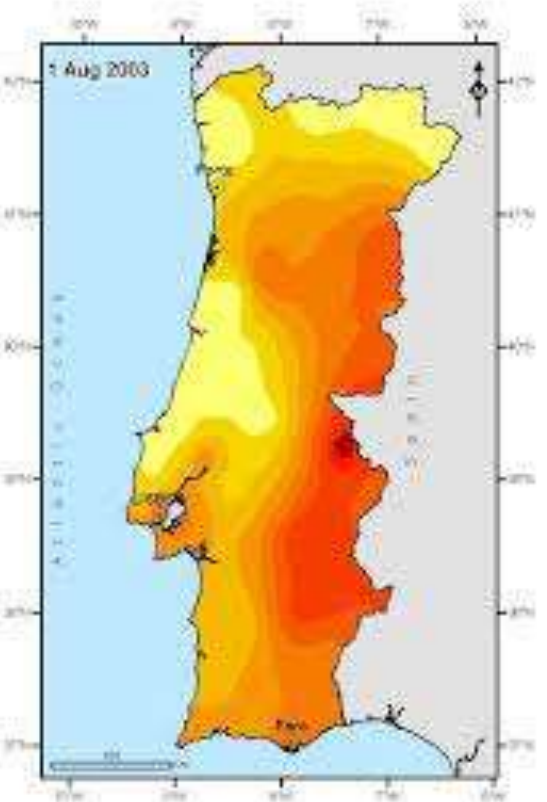
Tipos de dados

- Nominais;
- Categorias:
  - Atribui categorias a objetos:
    - Podem ser valores numéricos, mas são **não ordenados**;
    - Código postal;
    - Sexo;
    - Cor dos olhos;
- Ordinais;
- Intervalos;
- Rácios.

ISLab
Synthetic Intelligence Lab

Tipos de dados

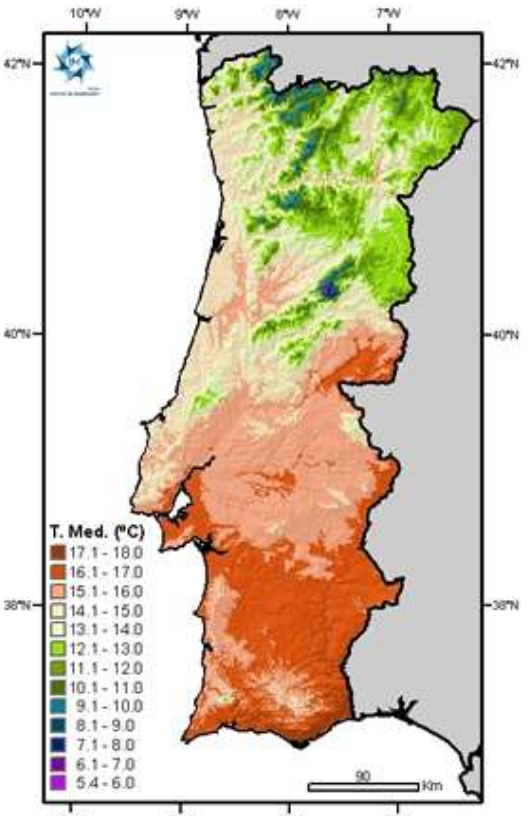
- Nominais;
- Categorias;
- Ordinais:
  - Os valores podem ser ordenados naturalmente;
    - Classificação: excelente, bom, suficiente, etc.;
    - Temperatura: frio, morno, quente;
- Intervalos;
- Rácios.



ISLab
Synthetic Intelligence Lab

Tipos de dados

- Nominais;
- Categorias;
- Ordinais;
- Intervalos:
  - É possível calcular a distância entre dois valores;
    - Temperatura;
    - Humidade;
- Rácios.



ISLab
Synthetic Intelligence Lab

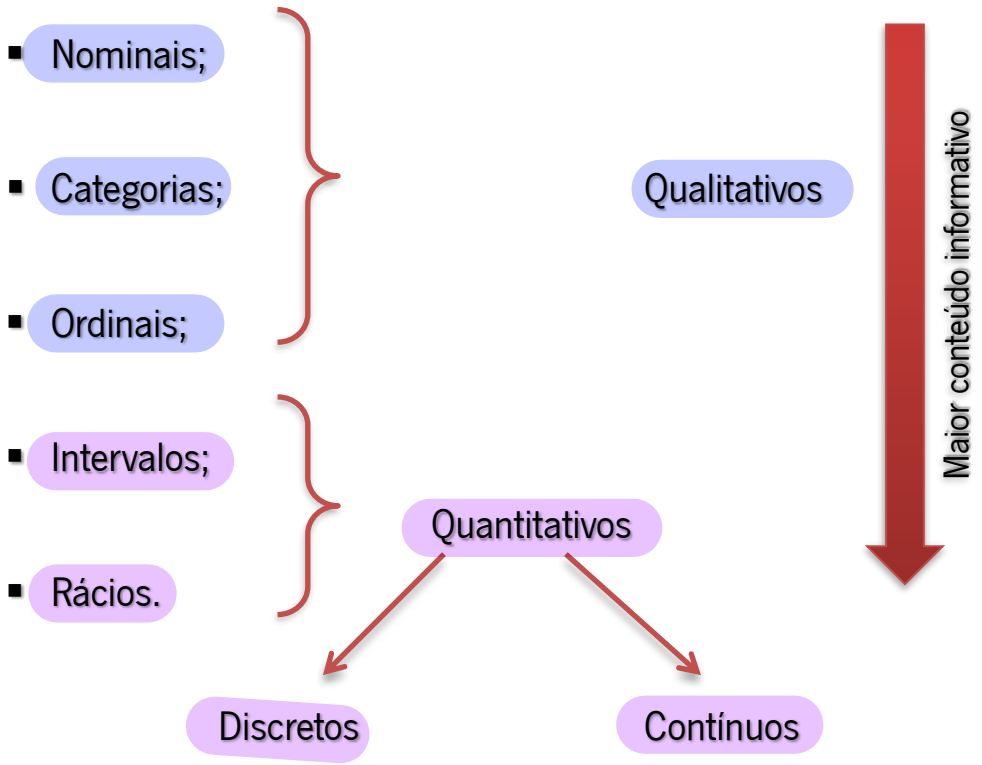
Tipos de dados

- Nominais;
- Categorias;
- Ordinais;
- Intervalos;
- Rácios:
  - Os valores podem ser utilizados para determinar um rácio significativo entre eles:
    - Salário;
    - Balanço bancário.

Country	Converted values			National rates and developments		
	2020 (€)	2021 (€)	Change (%)	2020 (€ unless otherwise stated)	2021 (€ unless otherwise stated)	Change (%)
Luxembourg	2,142	2,202	2.8	2,142/month	2,202/month	2.8
UK*†	1,790	1,903	6.3	GBP 8.72/hour	GBP 8.91/hour	2.2
Ireland*	1,707	1,724	1.0	10.1/hour	10.2/hour	1.0
Netherlands	1,654	1,685	1.9	1,654/month	1,685/month	1.9
Belgium*	1,626	1,626	0.0	1,626/month	1,626/month	0.0
Germany	1,584	1,610	1.6	9.35/hour	9.5/hour	1.6
France	1,539	1,555	1.0	1,539/month	1,555/month	1.0
Slovenia	1,019	1,110	8.9	1,019/month	1,110/month	8.9
Spain	1,108	1,108	0.0	1,108/month	1,108/month	0.0
Malta	777	785	1.0	179/week	181/week	1.0
Portugal	741	776	4.7	741/month	776/month	4.7
Greece	758	758	0.0	758/month	758/month	0.0
Lithuania	607	642	5.8	607/month	642/month	5.8
Slovakia	580	623	7.4	580/month	623/month	7.4
Poland	611	614	0.5	PLN 2,600/month	PLN 2,800/month	7.7
Estonia	584	584	0.0	584/month	584/month	0.0
Czechia	575	579	0.8	CZK 14,600/month	CZK 15,200/month	4.1
Croatia	546	563	3.1	HRK 4,063/month	HRK 4,250/month	4.6
Latvia	430	500	16.3	430/month	500/month	16.3
Romania	466	472	1.3	RON 2,230/month	RON 2,300/month	3.1
Hungary†	487	467	-4.1	HUF 161,000/month	HUF 167,400/month	4.0
Bulgaria	312	332	6.6	BGN 610/month	BGN 650/month	6.2

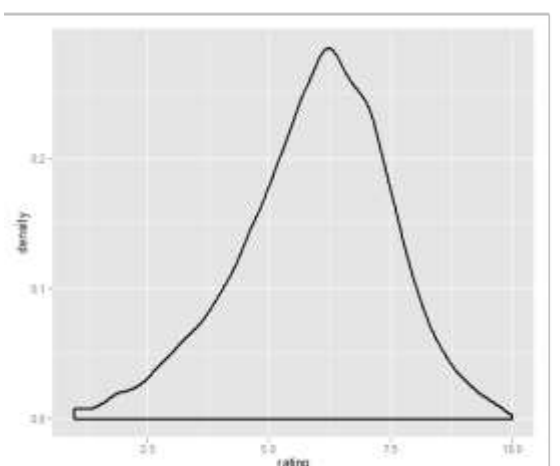
ISLab
Synthetic Intelligence Lab

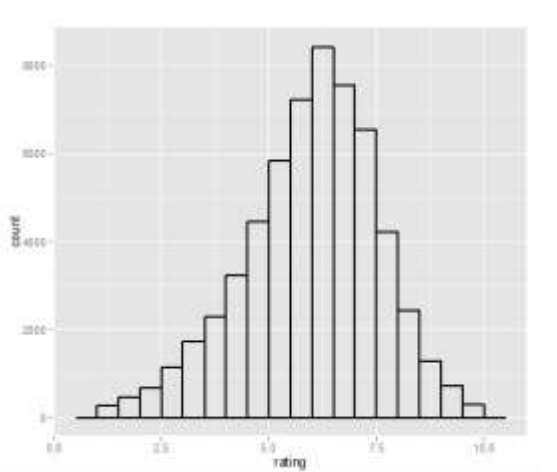
Tipos de dados



Discretização/Enumeração

- Utiliza-se a discretização (ou enumeração) para reduzir o número de valores de um atributo contínuo, dividindo-o em intervalos;
  - Os métodos mais utilizados (Naïve Bayes, CHAID, etc.), requerem valores discretos;
  - Redução do tamanho dos dados;
  - Método utilizado para produzir sumariação dos dados;
  - (Sinónimo de *binning*.)

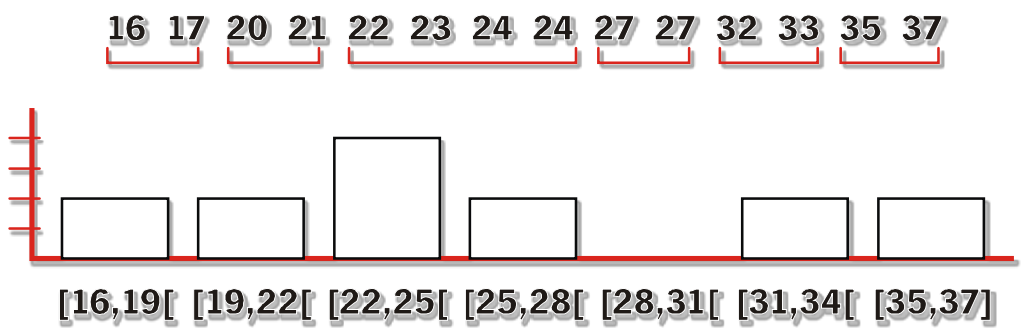




ISLab
Synthetic Intelligence Lab

Discretização de igual largura

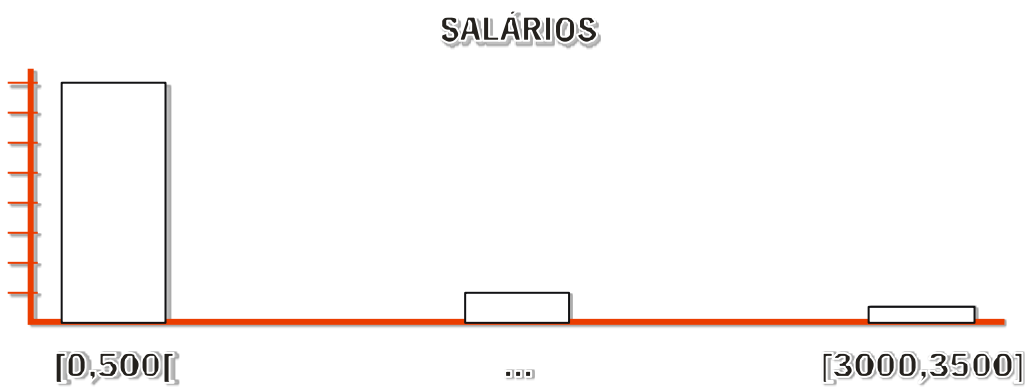
- *Equal-width binning*:
- Divide a gama de valores em N intervalos de igual largura, resultando numa grelha uniforme;
- Sendo A e B os limites da gama de valores, a largura dos intervalos será L = (B - A) / N:



ISLab
Synthetic Intelligence Lab

Discretização de igual largura

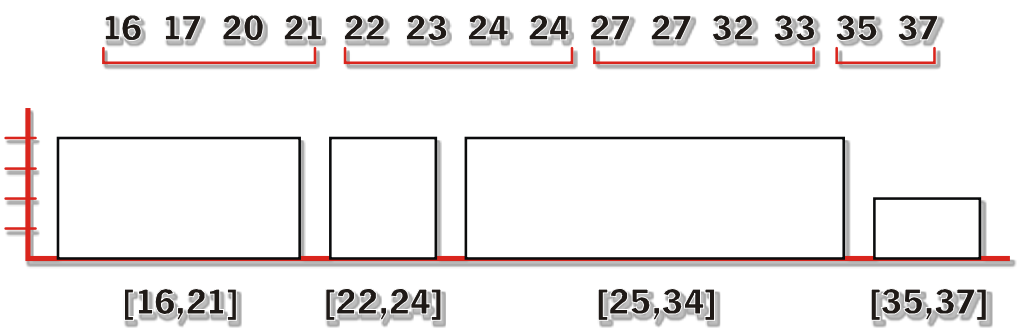
- Vantagens:
  - Simples e fácil de implementar;
  - Produz abstrações de dados razoáveis;
  - Não supervisionado;
  - Quem determina N?;
  - Sensível a valores fronteira.



- Desvantagens:

Discretização de igual altura

- *Equal-height binning*:
- Divide a gama de valores em N intervalos, contendo, cada um, **aproximadamente** a mesma quantidade de valores:



ISLab
Synthetic Intelligence Lab

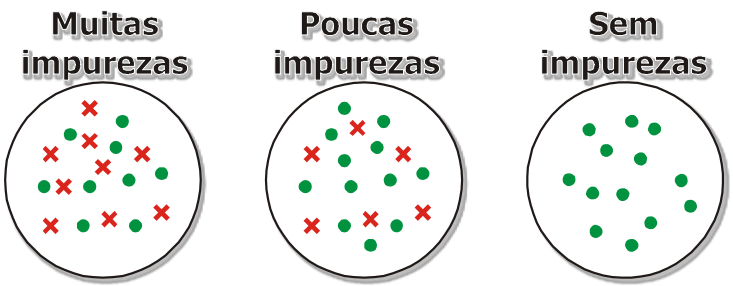
Discretização de igual altura

- Normalmente preferida à discretização de igual largura, uma vez que permite evitar o "amontoar" de valores;
- Na prática, utiliza-se uma discretização de "quase-igual" altura, garantindo intervalos mais intuitivos;
- Deverá impedir a dispersão de valores frequentes por diferentes intervalos;
- Deverá criar intervalos separados para valores especiais ("0").

ISLab
Synthetic Intelligence Lab

Discretização: outros métodos

- Método 1R:
  - Método supervisionado, baseado na divisão por *binning*;
- Discretização baseada em Entropia;
- Discretização baseada em Impurezas;



- Deteção de limites;
- etc.

ISLab
Synthetic Intelligence Lab

Limpeza de dados

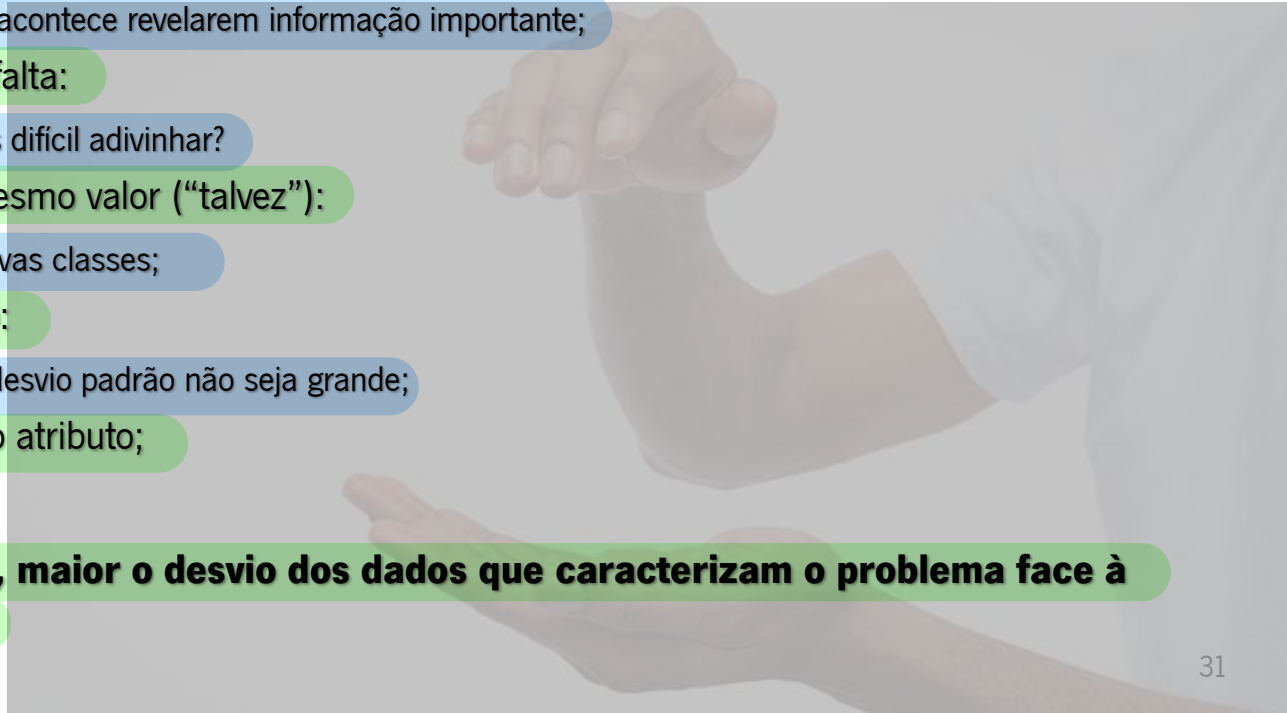
- Ausência de valores em determinados atributos devido a:
  - inconsistência;
  - dados não registados;
  - análise incorreta;
  - dados registados de forma errada;
  - etc.

- **A ausência de dados pode revelar algo sobre que campos não foram preenchidos!**

ISLab
Synthetic Intelligence Lab

Limpeza de dados: como tratar a ausência de dados?

- Ignorar os registos onde faltam os dados e lidar, apenas com os dados conhecidos;
  - não aconselhável se a quantidade de dados em falta em cada atributo for elevada;
- Ignorar os atributos onde faltam os dados;
  - não aconselhável se os atributos onde acontece revelarem informação importante;
- Preencher (manualmente) os dados em falta:
  - é mais trabalhoso preencher ou é mais difícil adivinhar?
- Preencher os dados em falta com um mesmo valor ("talvez"):
  - pode criar tendências nos dados ou novas classes;
- Preencher com o valor médio do atributo:
  - pouco impacto negativo, desde que o desvio padrão não seja grande;
- Preencher com o valor mais frequente do atributo;

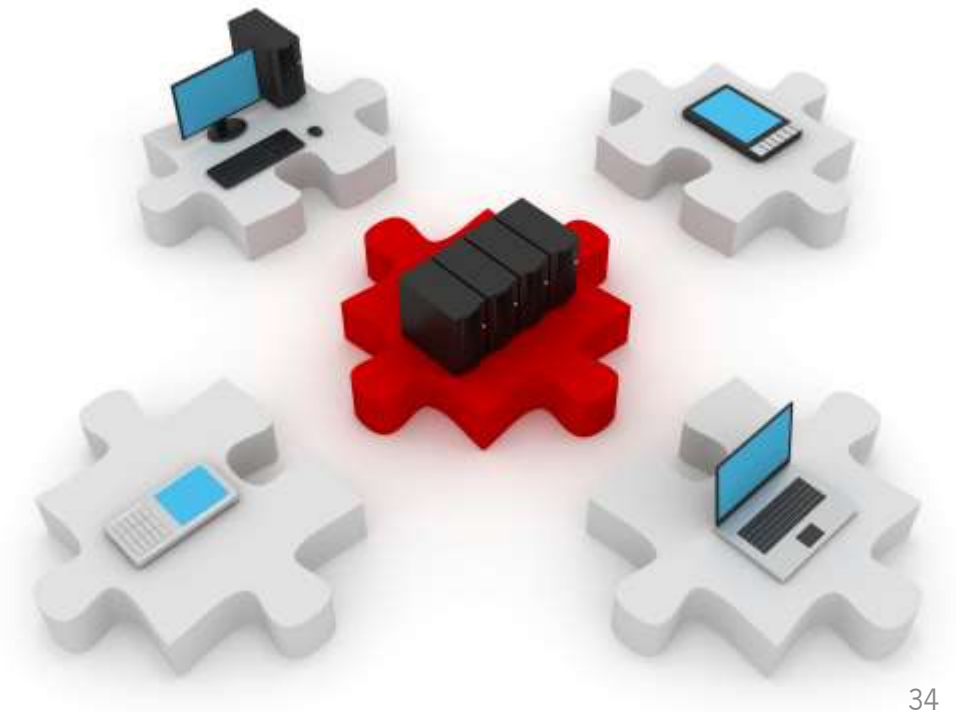


- **Quantos mais valores "inventados", maior o desvio dos dados que caracterizam o problema face à realidade que o problema ilustra!**

ISLab
Synthetic Intelligence Lab

Integração de dados

- Os dados que caracterizam o problema podem ter proveniências diversas;
- O objetivo da integração é o de compor um conjunto de peças de informação numa coleção coerente e integrada de dados.
- Detetar e resolver conflitos entre os dados:
  - qual a fonte de dados mas fiável, quando os valores que transportam são inconsistentes?
- Integração exige "**conhecimento do negócio**".



Transformação de dados

- Alisamento (*smoothing*):
  - Remover lixo/ruído dos dados (*binning*, regressão, *clustering*);
- Agregação;
- Generalização;
- Construção de Atributos;
- Uniformização;
- Deteção de valores atípicos.

Transformação de dados

- Alisamento (*smoothing*);
- Agregação:
  - Pressupõe que o resultado sumaria os dados iniciais;
    (resumo de vendas trimestrais, durante 5 anos, em valores anuais)
- Generalização;
- Construção de Atributos;
- Uniformização;
- Deteção de valores atípicos.

ISLab
Synthetic Intelligence Lab

Transformação de dados

- Alisamento (*smoothing*);
- Agregação;
- Generalização:
  - Hierarquização de conceitos:
    - distrito → cidade → rua;
    - Valores diferentes: 18 → centenas → (largos) milhares
- Construção de Atributos;
- Uniformização;
- Deteção de valores atípicos.

ISLab
Synthetic Intelligence Lab

Transformação de dados

- Alisamento (*smoothing*);
- Agregação;
- Generalização;
- Construção de atributos:
  - Construção de novos atributos a partir de outros
    (cálculo do preço líquido baseado no preço ilíquido e no IVA);
- Uniformização;
- Deteção de valores atípicos.

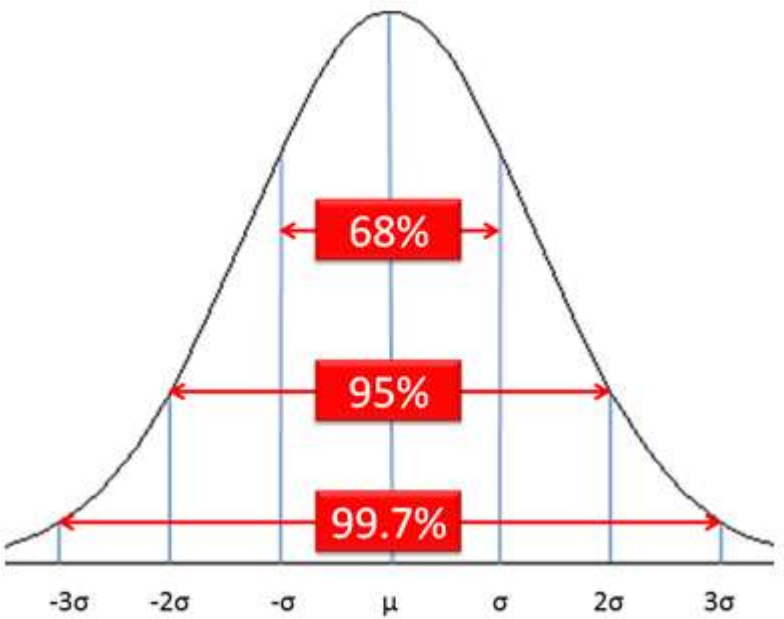
Transformação de dados

- Alisamento (*smoothing*);
- Agregação;
- Generalização;
- Construção de atributos;
- Uniformização:
  - Pretende evitar que atributos com uma gama alargada de valores sobressaiam em relação a outros atributos com menor quantidade de valores:
    - Normalização (*normalization*: [0;1]);
    - Padronização (*standardization/Z-score normalization*: $\bar{x}$=0; ∂=1);
- Deteção de valores atípicos.

ISLab
Synthetic Intelligence Lab

Transformação de dados

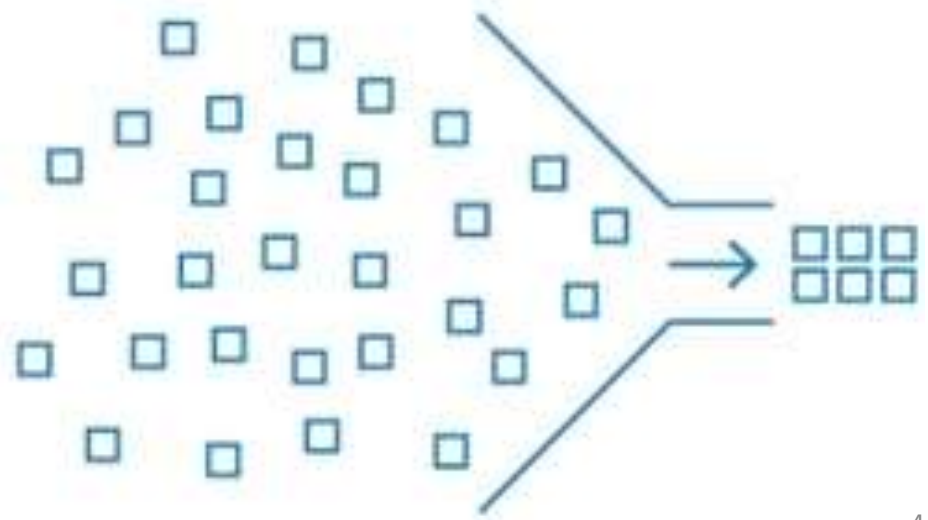
- Alisamento (*smoothing*);
- Agregação;
- Generalização;
- Construção de atributos;
- Uniformização;
- Deteção de valores atípicos:
  - Por visualização:
    - Box plots
    - Z-Score (desvio padrão)



https://sixsigmastudyguide.com/z-scores-z-table-z-transformations/

Redução de dados

- Um *Data Warehouse* pode armazenar largos terabytes de dados;
- Realizar tarefas de EC em tais quantidades de dados pode tornar-se impraticável!
- A **Redução de dados** pretende obter uma representação reduzida do volume de dados, mas produzindo os mesmos (ou quase os mesmos) resultados analíticos.



ISLab
Synthetic Intelligence Lab

Redução de dados: estratégias

- Construção de cubos de dados:
  - as operações de agregação são aplicadas de modo a construir cubos de dados;
- Redução de dimensões:
  - remoção de atributos que se mostrem irrelevantes, redundantes ou pouco interessantes para a análise;
  - *Principle Component Analysis* (PCA);
- Compressão de dados:
  - aplicação de técnicas de compressão ou de transformação para comprimir a representação dos dados originais;
- Redução de quantidade:
  - redução do volume de dados (técnicas paramétricas ou não paramétricas);
- Discretização e generalização de conceitos:
  - redução da quantidade de valores por atributo.