UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Visualização de distribuições de dados complexos provenientes de fontes heterogêneas

Natalia de Fatima Martins

Dissertação de Mestrado do Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-CCMC)

ICMC USP
SÃO CARLOS

Natalia de Fatima Martins

Visualização de distribuições de dados complexos provenientes de fontes heterogêneas

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Mestra em Ciências – Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientador: Prof. Dr. Caetano Traina Junior

USP – São Carlos
Maio de 2023

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

M379v

Martins, Natalia de Fatima
Visualização de distribuições de dados complexos provenientes de fontes heterogêneas / Natalia de Fatima Martins; orientador Caetano Traina Junior. -- São Carlos, 2023.
83 p.

Dissertação (Mestrado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2023.

1. Visualização de Dados. 2. Eletronic Health Records. 3. OMOP CDM. 4. Interoperabilidade. 5. Dados heterogeneos. I. Traina Junior, Caetano , orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Natalia de Fatima Martins

Visualization of complex data distributions from heterogeneous sources

Dissertation submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP – in accordance with the requirements of the Computer and Mathematical Sciences Graduate Program, for the degree of Master in Science. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Prof. Dr. Caetano Traina Junior

USP – São Carlos
May 2023

Este trabalho é dedicado aos meus pais, Silvia de F. M. Martins e José N. Martins que, desde muito antes da conclusão deste mestrado, se dedicaram e abdicaram de si mesmos em muitos momentos para me oferecerem um caminho cheio de oportunidades.

Ao meu namorado, noivo e agora, marido, Dr. Hendrik Marques Soares, que me apoiou de todas as formas possíveis nos mais de mil dias de dedicação à pesquisa.

Ao melhor cientista e ser humano que eu poderia ter tido como orientador, Prof. Dr. Caetano Traina Júnior, excepcional em sua capacidade de ouvir, ensinar e de representar o verdadeiro espírito da ciência.

Por fim, para minha querida avó, Sylvia Marino (in memoriam), que foi exemplo de força e determinação.

Obrigada, eu não teria conseguido sem vocês.

AGRADECIMENTOS

Agradeço a Deus por ter me guiado e amparado nos momentos desafiadores deste trabalho e desta pandemia. Agradeço aos meus amigos de laboratório, Afonso, Alex, Alexis e em especial ao agora mestre João Victor de Oliveira Novaes pelo companheirismo e pelos momentos juntos, em que rimos e compartilhamos vitórias e naqueles em que me deram apoio e vieram em meu auxílio. Agradeço também aos colegas de pesquisa, em especial ao Daniel Mario de Lima, que sempre me auxiliou prontamente e contribuiu muito para que eu pudesse realizar este mestrado. Agradeço à todo o grupo MiVisBD do ICMC, especialmente à Profa Dra. Agma Juci Machado Traina que também sempre esteve disposta à ouvir e a auxiliar no que fosse necessário. Agradeço à Universidade de São Paulo (USP), a FAPESP, CAPES e CNPQ pelos recursos disponibilizados, fundamentais para a conclusão deste mestrado. Finalmente, aos meus amigos, que foram pacientes enquanto eu os trocava aos finais de semana por Bancos de Dados.

“Em algum lugar, alguma coisa incrível
está esperando para ser descoberta.”
(Carl Sagan)

RESUMO

MARTINS, N. F. **Visualização de distribuições de dados complexos provenientes de fontes heterogêneas**. 2023. 83 p. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2023.

Os Sistemas de Gerenciamento de Bases de Dados (SGBD) baseados na Teoria Relacional foram desenvolvidos para atender às necessidades de armazenagem e recuperação de dados representados por valores numéricos, datas e pequenas cadeias de caracteres, chamados genericamente de dados escalares. Consultas sobre dados escalares são feitas por comparações baseadas em relações de identidade ou relações de ordem. Com a evolução da tecnologia da informação, se faz necessário organizar, armazenar e recuperar outros tipos de dados, tais como imagens, vídeos, séries temporais, sequências genômicas etc. A esses, nos referimos como dados complexos, porque eles não são comparados diretamente, mas por meio de funções de extração de características aplicadas sobre os dados. Portanto, uma vez extraídas, as características são utilizadas no lugar dos dados originais para executar as comparações.

Considerando, ainda, a crescente e genuína necessidade de se analisar grandes volumes de dados de diferentes e diversas áreas do conhecimento, em destaque, a área médica, objeto deste trabalho e avaliando as inerentes limitações das técnicas mais frequentes, apresentamos, neste trabalho, uma abordagem alternativa, utilizando ferramentas de visualização (Tableau) para avaliar e gerar *insights* sobre objetos complexos (pacientes) através da modelagem de EHR (Eletronic Health Records) disponibilizados pelo Hospital do Coração (InCor). Estes, por sua vez, estão armazenados sobre o padrão OMOP, uma arquitetura de dados Entidade-valor-Atributo que também é desafiadora para analistas e outros profissionais da área de dados. O resultado deste trabalho é apresentar um caminho prático e aplicável que oferece benefícios potenciais tanto para técnicas de avaliação de objetos complexos quanto para descoberta de conhecimento na área médica.

Palavras-chave: Visualização de Dados, Dados Heterogêneos, Interoperabilidade, Registros médicos EHR, OMOP-CDM.

ABSTRACT

MARTINS, N. F. **Visualization of complex data distributions from heterogeneous sources**. 2023. 83 p. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2023.

Database Management Systems (DBMS) based on Relational Theory were developed to meet the needs of storing and retrieving data represented by numeric values, dates, and short strings of characters, generically called scalar data. Queries on scalar data are made by comparisons based on identity relations or order relations. With the evolution of information technology, it has become necessary to organize, store and retrieve other types of data, such as images, videos, time series, genomic sequences, etc. These we refer to as complex data, because they are not compared directly, but by means of feature extraction functions applied on the data. So once extracted, the features are used in place of the original data to perform the comparisons.

the growing and genuine need to analyze large volumes of data from different and diverse areas of knowledge, in particular, the medical area, the object of this work

Considering, furthermore, the growing and genuine need to analyze large volumes of data from different and diverse areas of knowledge, in particular, the medical area, object of this work and evaluating the inherent limitations of the most frequent techniques, we present in this paper an alternative approach, using visualization tools (Tableau) to evaluate and generate insights about complex objects (patients) through the modeling of EHR (Electronic Health Records) made available by the Hospital do Coração (InCor). These, in turn, are stored over the OMOP standard, an Entity-Value-Attribute data architecture that is also challenging for analysts and other data professionals. The result of this work is to present a practical and applicable way forward that offers potential benefits for both complex object evaluation techniques and knowledge discovery in the medical field.

Keywords: 1. Data Visualization 2. Heterogeneous Data 3. Interoperability. 4. Eletronic Health Records EHR 5. OMOP CDM.

LISTA DE ILUSTRAÇÕES

LISTA DE CÓDIGOS-FONTE

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

1.1 Panorama e Conceitos

Os Sistemas de Gerenciamento de Bases de Dados (SGBD) baseados na Teoria Relacional começaram a ser desenvolvidos no final da década de 1960 (CODD, 1970) para atender às necessidades de armazenagem e recuperação de dados representados como valores numéricos, datas e pequenas cadeias de caracteres, chamados genericamente "dados escalares" (DATE, 2009). Desde então, o avanço desses sistemas aplicados ao desenvolvimento de softwares os colocaram presentes em praticamente todas as áreas da atividade humana, sendo, portanto, a opção por excelência para o armazenamento e consultas aos dados. A recuperação de dados escalares é feita comparando-se os elementos usando operadores baseados em Relações de Igualdade (RI, expressos pelos símbolos $=$ e $\neq$) ou em Relações de Ordem (RO, expressos por $<$, $\leq$, $>$ e $\geq$), constituindo os operadores comumente chamados "*the big six*" do modelo relacional (MELTON; SIMON, 2002).

Apesar da ampla aceitação desses sistemas, a inevitável e acelerada evolução da tecnologia da informação traz consigo a necessidade de se armazenar e recuperar outros tipos de dados, tais como multimídia (imagens, áudio e vídeo), informações geo-referenciadas, dados multidimensionais, sequências, séries temporais, etc. A esses dados nos referimos como "dados complexos", porque a comparação entre eles não é feita diretamente, mas usando inúmeras funções de comparação, que podem variar de acordo com a necessidade das aplicações, já que os objetos possuem muitos *aspectos* pelos quais se pode querer comparar. Evidentemente, comparações baseadas em relações de igualdade têm pouca utilidade em consultas sobre dados complexos, pois é raro que existam dois objetos complexos – por exemplo, duas imagens – exatamente iguais. As comparações baseadas em relações de ordem avaliam a precedência de um objeto sobre outro e geralmente também não são aplicáveis, pois os domínios onde dados complexos são representados usualmente não atendem a essa propriedade (BRUNO; CHAUDHURI, 2010). Um exemplo bastante ilustrativo é considerar a comparação entre a imagem de um carro

vermelho e de uma rosa, também vermelha: não tem sentido utilizar os operadores baseados em ordenação para essa tarefa. Um outro exemplo de interesse para este trabalho são objetos associados ao tratamento da saúde, como exames, drogas e procedimentos: também não tem sentido comparar esses elementos por ordem nem igualdade.

Uma abordagem interessante é permitir que cada aspecto do objeto tratado possa ser usado em operações de comparação e, para isso, é comum que sejam usadas funções que extraem diferentes "características" dos dados, as quais são usadas ao invés do dado original nas comparações. Para distinguir, neste trabalho usamos o termo "objeto" para referir ao dado complexo original (a imagem, o áudio, etc.) e o termo "elemento" para referir às características extraídas (usadas em geral como elemento de comparação).

Potencialmente, cada tipo de dado complexo tem diversos aspectos que podem ser de interesse nas comparações, tais como cor, textura ou forma das imagens, ritmo ou entonação em áudio, etc. Além disso, um mesmo aspecto pode ser comparado extraindo-se as características correspondentes de diversas maneiras diferentes, tais como, especificamente para cor de imagens, as distribuições locais ou global das cores, os diferentes modelos de representação de cor, etc.

1.1.1 Objetivos Gerais do Projeto

Uma comparação feita indiretamente usando características extraídas automaticamente dos objetos complexos faz com que, frequentemente, diversos objetos cujas características sob um determinado aspecto são semelhantes não sejam de fato semelhantes.

Por exemplo, duas fotos com cores semelhantes podem ter conteúdo completamente diferentes. Assim, é frequente combinar-se diversos aspectos numa mesma consulta, visando reduzir a quantidade de falsos positivos (seletividade), tal como recuperar imagens com cor e forma semelhantes. Por exemplo, a seletividade da seleção de fotos escolhidas pelo aspecto 'forma' de um automóvel Ferrari usando o aspecto 'cor' vermelha deve ser pequena, embora as características de forma e cor tenham pouca correlação em outros pontos de busca.

Quando consideramos um contexto médico, é notável a existência de diversos tipos de objetos complexos, como por exemplo os pacientes, os procedimentos a que eles são submetidos, as instituições de saúde, os exames efetuados, as drogas utilizadas e muitos outros. Os dados escalares são representados tanto por atributos tradicionais, tais como o nome e a idade dos pacientes, quanto por atributos complexos. Para distinguir dentre ambos, nesta monografia, usamos o símbolo S para designar um atributo complexo e o símbolo S^{fx} para designar suas respectivas características extraídas. Por sua vez, as características tanto podem ser denotativas – as características que podem ser extraídas por processos automáticos, como por exemplo, a extração de características de imagens por algoritmos de processamento de imagens; quanto conotativas – aquelas são obtidas externamente ao objeto, como por exemplo, usando pessoas para rotular uma imagem com *tags* ou efetuando "medidas" em um paciente por meio de exames

médicos.

A existência de possíveis correlações entre os espaços de consulta é um fator que afeta fortemente a seletividade resultante de uma conjunção de condições, tornando a escolha de características adequadas ainda mais importante. De fato, se forem escolhidos aspectos correlacionados, a capacidade de filtragem (seletividade) de um aspecto sobre os elementos selecionados por outro pode ser pequena. Mais ainda, o índice de correlação entre os dois aspectos pode variar dependendo das regiões do espaço onde é feita a busca.

Em termos de representação de consultas, isso é feito expressando-se a conjunção (executada como a intersecção) das seleções de objetos que atendem a cada critério individual, o que tanto tem um custo elevado de execução (deve-se processar as diversas operações de seleção e a intersecção dos resultados intermediários) quanto requer um conhecimento profundo por parte do usuário de quais são os aspectos que se sobrepõem o menos possível.

Considerando as dificuldades para expressar uma consulta que envolve comparações entre dados complexos, este projeto visa estudar **"Visualização de Distribuições de Dados Complexos Provenientes de Fontes Heterogêneas"**, visando facilitar a navegação através dos espaços de busca gerados pelos diversos aspectos usados para realizar as comparações sobre objetos complexos utilizando técnicas de visualização de dados.

A base de dados utilizada nesses trabalho é disponibilizada pelo Instituto do Coração de São Paulo (InCor) para o Projeto de Pesquisa Temático MiVisBD[1], do qual este projeto de mestrado é parte. Os dados estão sendo armazenados seguindo o padrão OMOP-CDM, que será apresentado mais a frente neste trabalho.

[1] Projeto Temático de Pesquisa FAPESP Nº2016/16361-0

CAPÍTULO

2

CONCEITOS

2.1 Conceitos Fundamentais

Nesta seção sumarizam-se os conceitos mais importantes para o desenvolvimento deste projeto, incluindo uma rápida discussão de como coleções de dados complexos são processadas para permitir que sejam comparadas por seu conteúdo em buscas por similaridade e uma descrição dos operadores de busca por similaridade estudados com mais frequência na literatura.

2.1.1 Consultas por similaridade em conjuntos de dados complexos

Seja $\mathbb{S}$ um domínio de dados complexos, $S \subseteq \mathbb{S}$ um conjunto de dados complexos desse domínio que está armazenado num atributo S e seja $s \in \mathbb{S}$ um objeto complexo. A busca em coleções de dados complexos em geral não compara diretamente os valores complexos originais, mas descrições e características que os descrevam ou representem. Assim, é necessário associar a cada objeto complexo s um ou mais características que o represente. Seja S^{f_x} um atributo que armazena a descrição dos elementos de S segundo um determinado aspecto f_x. Então, cada objeto complexo s $s_i \in \mathsf{S}$ que deve ser comparado segundo um aspecto f_x, deve ter uma descrição s^{f_x} associada.

Existem duas técnicas para associar atributos que representem as características dos objetos complexos. A primeira é chamada "recuperação por **descrição**", como por exemplo em "Recuperação de Imagens por Descrição" (DBIR na sigla em inglês – *Description-Based Image Retrieval*) (WU; JIN; JAIN, 2013). Ela corresponde à associação de atributos ao objeto complexo, em geral de tipo textual, cujos valores são obtidos externamente, e pode ser considerada uma **descrição conotativa** do objeto. Exemplos dessa técnica aplicada à coleções de imagens são: a associação de palavras-chave (*tags*) e descrições feitas por pessoas, tais como laudos em imagens de exames médicos (SIMPSON *et al.*, 2010); o uso dos metadados EXIF em fotografias; a associação automática de textos próximos às imagens em páginas da *web*; e a obtenção de dados

adicionais que auxiliem a descrição do objeto complexo por processos externos.

Um padrão representativo para o compartilhamento de descrições de imagens visando operações de busca é o JPSearch (ISO/IEC 24800) (DÖLLER *et al.*, 2013) administrado pelo *Joint Photographic Experts Group* (JPEG), que é baseado em descrições padronizadas das imagens em XML, baseado em ontologias de termos vinculados a domínios específicos de imagens. Descrições contribuem representando um significado mais contextual e interpretativo do objeto, por isso são também chamadas **descrições conotativas**, e a busca por descrição é chamada **"recuperação por contexto"**.

A segunda técnica para associar características aos objetos complexos é chamada "recuperação por **conteúdo**", como, por exemplo, em "Recuperação de Imagens por Conteúdo" (CBIR na sigla em inglês – *Content-Based Image Retrieval*) (AGGARWAL *et al.*, 2019; BRANCATI; CAMASTRA, 2016; WELTER *et al.*, 2012). Ela corresponde à associação de atributos obtidos a partir do próprio elemento complexo, extraídos por algoritmos de processamento automáticos chamados "**vetores de características**". Vetores de características podem tanto seguir o sentido matemático do termo, quando todas as características são de um mesmo tipo numérico e mensuradas na mesma unidade – como é o caso dos histogramas de cor ou textura de imagens (SAJJAD *et al.*, 2018) – como também podem ser sequências de medidas multivariadas, envolvendo medidas diferentes para diversas características e/ou com quantidades diferentes de medidas – como é o caso das poligonais que descrevem formas encontradas nas imagens (LIU *et al.*, 2007; TRAINA *et al.*, 2010). Características são também chamadas **descrições denotativas**.

Por simplicidade, quando não explicitamente indicado neste texto, usamos o termo **característica** para nos referirmos a uma "descrição denotativa" e o termo **descrição** restrito ao significado de "descrição conotativa". Usamos também indistintamente a notação $s^{f_x} = f_x(s_i)$ para representar a extração automática de uma característica ou a associação externa de uma descrição conotativa associada a um objeto s_i.

Em ambas as técnicas de associação de atributos, as comparações envolvendo atributos complexos são executadas sobre as descrições ou sobre os vetores de características correspondentes em lugar dos objetos complexos propriamente ditos. Cada comparação é feita por uma **função de atribuição** (ou de medida) **de similaridade**, das quais existem duas grandes classes estudadas na literatura (HAN; KAMBER; PEI, 2012):

a) **Coeficientes de Similaridade** d_{coef}: quanto maior o resultado de d_{coef}, maior é a similaridade entre o par de objetos comparados – exemplos são o coeficiente de Jaccard e o coeficiente do cosseno; e

b) **Funções de Distância** f_d: quanto maior o resultado de f_d, menor é a similaridade entre os objetos, sendo que a distância de um objeto para ele mesmo é zero – exemplos são a função Euclidiana para vetores e a função de distância de edição Levenshtein para cadeias de caracteres.

Diferentes áreas de pesquisa na Computação dão preferência ao uso de coeficientes de similaridade ou a funções de distância e, com frequência, é possível substituir uma pela outra (DEZA; DEZA, 2016). Por exemplo, o Coeficiente de Jaccard $c_{Jaccard}$ pode ser transformado numa "distância de Jaccard" $d_{Jaccard}$ fazendo $d_{Jaccard} = 1 - c_{Jaccard}$. A área de Bases de Dados em geral usa funções de distância f_d .

2.1.2 Espaços métricos

Do ponto de vista dos algoritmos de busca em grandes coleções de dados, é importante que cada função de atribuição de similaridade garanta propriedades que permitam usar técnicas de indexação para agilizar as buscas no espaço usado para representar os dados (TRAINA *et al.*, 2009; GÜLD *et al.*, 2007). Nesse sentido, a existência de uma referência para comparar um elemento com ele mesmo nas funções de distância ($f_d(s_i, s_i) = 0$) pode ser uma propriedade interessante, não compartilhada genericamente pelos coeficientes de similaridade. Outras propriedades interessantes são obtidas quando as descrições e os vetores de características são representados em espaços métricos. Assim, no lugar de propriedades baseadas em relações de identidade ou de ordem, usam-se as propriedades dos espaços métricos para construir estruturas de acesso aos dados (CHEN *et al.*, 2015).

Um espaço métrico é definido como um par $\langle \mathbb{S}, f_d \rangle$, onde $\mathbb{S}$ é o conjunto dos elementos que atendem aos requisitos do espaço, chamado **domínio** dos dados, e $f_d : \mathbb{S} \times \mathbb{S} \to \mathbb{R}^+$ é uma função de distância, chamada também de **métrica**, que atende às quatro propriedades (LIMA, 1993) seguintes, para $\forall s_1, s_2, s_3 \in \mathbb{S}$.

- **não-negatividade**: $0 < f_d(s_1, s_2) < \infty$ se $s_1 \neq s_2$
- **identidade dos indiscerníveis**: $f_d(s_1, s_1) = 0$;
- **simetria**: $f_d(s_1, s_2) = f_d(s_2, s_1)$; e
- **desigualdade triangular**: $f_d(s_1, s_2) \leq f_d(s_1, s_3) + f_d(s_3, s_2)$.

Um conjunto de dados S armazenado em um atributo S em uma relação qualquer é dito estar num espaço métrico quando $S \subset \mathbb{S}$. A função de distância é usada para quantificar o quão similar são dois elementos, o que habilita executar consultas por similaridade baseadas nessa função, assumindo que quanto menor é a distância, mais similares os dois elementos são.

Funções de distância

As funções de distâncias são definidas segundo o tipo de dados do espaço $\mathbb{S}$. Os tipos de dados mais comuns usados em comparações por similaridade são dados dimensionais, sequências e conjuntos, mas outros tipos são usados também, tais como séries temporais, hierarquias e grafos (NIEDERMAYER, 2015). É importante lembrar que, em aplicações práticas, nem todas as funções de distâncias usadas respeitam as propriedades de uma métrica, mas apenas as que

respeitam podem ser usadas para construir estruturas de acesso que agilizam a execução das consultas, chamadas Métodos de Acesso Métrico (MAM).

Descrições $f_x(s_i)$ extraídas automaticamente dos objetos complexos s_i em geral têm a estrutura de **vetores com dimensionalidade definida** e são chamados ” vetores de características”. Assim, todas as instâncias têm as mesmas dimensões (os mesmos atributos), embora não necessariamente todos tenham a mesma escala e unidade de medida. Quando todas as dimensões têm a mesma unidade de medida (por exemplo, histogramas de cor dos *pixels* de uma imagem), é comum usar funções da família *Minkovsky*, tais como as distâncias Chebychev, Euclidiana ou *Manhattan*. Quando as dimensões correspondem a unidades de medida distintas ou têm domínios variados, como é o caso por exemplo das estatísticas calculadas sobre matrizes de coocorrência de cores numa imagem, é comum usar funções que ponderam as diversas dimensões, tais como as distâncias Euclidiana Padrão (Standardized Euclidean distance) e χ-quadrado (ANDERSON, 1984) Outro exemplo é a distância Mahalanobis, que considera a covariância entre dois subconjuntos de dimensões e representa a similaridade entre duas amostras (DEZA; DEZA, 2016).

Exemplos de **sequências** frequentemente comparadas por similaridade são os textos e as sequências genéticas. Funções de distância de edição tais como Levenshtein e Damerau-Levenshtein (DEZA; DEZA, 2016) são usadas para sequencias em geral, enquanto algumas funções foram desenvolvidas especificamente para determinados tipos de dados, como as funções de Jaro-Winkler (HERZOG; SCHEUREN; WINKLER, 2007) para endereços e funções baseadas nos padrões de alinhamento Blast e Fasta para sequencias genéticas (ALTSCHUL *et al.*, 1994; GUSFIELD, 1997) (que não atendem às propriedades de uma métrica). Distâncias para comparar **conjuntos** comuns incluem as distâncias de Jaccard e Steinhaus (DEZA; DEZA, 2016).

Devido à complexidade das distâncias para estruturas **hierárquicas e grafos**, às vezes são usadas distâncias baseadas em programação dinâmica, como por exemplo, para textos em XML (JAGADISH; MENDELZON; MILO, 1995; TEKLI *et al.*, 2011) e para **séries temporais** (PAPAPETROU *et al.*, 2011; KEOGH, 2002), as quais nem sempre atendem às propriedades métricas.

2.1.3 Atributos complexos

O tratamento e as buscas por similaridade sobre dados complexos têm sido estudados em diversas áreas de pesquisa e a terminologia usada por elas varia bastante. Neste projeto, adotamos a terminologia descrita a seguir. Seguindo a maneira usual da literatura de Bases de Dados, (ELMASRI; NAVATHE, 2011; DATE, 2009; GARCÍA-MOLINA; ULLMAN; WIDOW, 2011), qualquer atributo é representado pelo símbolo A, sendo seus valores a_i obtidos em um domínio $\mathbb{A}$. O conjunto de todos os valores de um domínio armazenados num atributo é representado como A, de tal maneira que $A \subseteq \mathbb{A}$ e $\forall a_i \in A \Rightarrow a_1 \in \mathbb{A}$ e A é um atributo amostrado em $\mathbb{A}$. Quando uma busca deve ser executada num atributo A, ela é feita no conjunto de valores efetivamente

armazenado e, portanto, o espaço de busca sobre um atributo A é o conjunto A.

Um atributo S é chamado de "**atributo complexo**" quando seus valores são obtidos de um domínio de dados complexo $\mathbb{S}$ e tem seus valores complexos s_i comparados usando uma função de distância associada a $\mathbb{S}$.

Uma função de distância compara os elementos $s_i \in \mathbb{S}$ de um domínio de dados complexos $\mathbb{S}$ *diretamente*, por exemplo atributos de tipo texto comparados pela função L_{edit}; ou *indiretamente* usando extratores de características, por exemplo atributos de tipo imagem dos quais se extraem histogramas de cor, que são comparados pela função *Manhattan*. O par $d = \langle f_x, f_d \rangle$ é chamado de "**descritor**", onde f_x é um extrator de característica ou uma função que obtém uma descrição conotativa associada, e f_d é uma função de distância. Usamos o símbolo d para indicar uma comparação de valores complexos segundo um descritor. Portanto, comparar segundo um descritor corresponde a extrair as características dos dois valores e a seguir usa-las para executar uma função de distância. Ou seja, dado um descritor $\mathbb{S} \mapsto d$, $d = \langle f_x, f_d \rangle$ associado a $\mathbb{S}$ e dois valores $s_1, s_2 \in \mathbb{S}$, $d(s_1, s_2) = f_d(f_x(s_1), f_x(s_2))$.

Para unificar o tratamento de comparações feitas direta ou indiretamente, assumimos que atributos comparados diretamente usam como característica um extrator "*neutro*" ($f_x = id \mid id(a) = a$), que corresponde ao próprio valor complexo. Para distinguir entre o valor complexo armazenado no atributo e suas respectivas características, usamos o termo "**objeto**" para indicar o valor complexo do atributo e "**elemento**" ao valor das características ou descrições, ou seja, o valor que é efetivamente comparado pela função de distância. Portanto, "elemento" e "objeto" são conceitos indistinguíveis quando se referem a atributos comparados diretamente, mas são distintos para atributos comparados indiretamente, sendo esta a situação mais frequente para atributos complexos.

Extrair características dos objetos complexos para então compará-los não altera o conceito fundamental de que, para agilizar a recuperação dos dados em uma base de dados, a função de comparação deve atender às propriedades de uma métrica. Cada extrator forma um espaço de características que deve ser associado a uma métrica, portanto cada descritor gera um espaço métrico (TRAINA *et al.*, 2010; SAMET, 2006) e, assim, cada descritor associado ao conjunto de valores armazenado em um atributo complexo constitui um **espaço de distâncias**, que também é um **espaço de busca** para o objeto complexo. Com funções de distância adequadas sobre as características extraídas dos objetos do domínio de dados complexo, os espaços métricos tornam-se excelente alternativa para executar consultas por similaridade com eficiência (POLA; TRAINA; JR, 2009; MALIK *et al.*, 2009).

2.1.4 Operadores de comparação, busca e seleção

Para aplicações em Bases de Dados, as funções de distância e os extratores de características são considerados uma "*caixa preta*", geralmente definida por um especialista no domínio

da aplicação. As características e descrições associadas aos objetos complexos são armazenadas junto a eles, de maneira que executar uma comparação sobre objetos armazenados frequentemente corresponde a executar apenas a função de distância sobre os dados associados. Portanto, uma vez definida uma função de distância adequada, de preferência que respeite as propriedades de uma métrica, os operadores de comparação por similaridade se aplicam a muitos tipos de dados complexos, incluindo dados multidimensionais, dados multimídia, textos longos, sequências, etc.

Uma comparação c entre valores de um domínio complexo $\mathbb{S}$ é representada como uma expressão $c = (s_i\ \theta(d)\ s_j)$, onde $s_i, s_j \in \mathbb{S}$ e $\theta(d)$ é um **operador de comparação por similaridade** válido em $\mathbb{S}$, que usa o descritor d. Existem dois operadores de comparação por similaridade fundamentais: o comparador por abrangência ("*similarity range*": $\theta(d) = Rng(d,\xi)$) e o comparador pelos k-vizinhos mais próximos ("*k-nearest neighbors*": $\theta(d) = k\text{-}NN(d,k)$). O operador de comparação por abrangência tem como parâmetros o descritor d e um raio máximo ξ. O resultado da comparação $(s_i\ Rng(d,\xi)\ s_j)$ sobre os objetos complexos s_i e s_j retorna `Verdade` sempre que $d(s_i,s_j) \leq \xi$. O operador de comparação pelos k-vizinhos mais próximos tem como parâmetros o descritor d e uma quantidade k. O resultado da comparação $(s_i\ k\text{-}NN(d,k)\ s_j)$ retorna `Verdade` sempre que s_j é um dos k vizinhos mais próximos a s_i no espaço de busca. Como ambas as comparações por similaridade precisam do parâmetro d, usamos a notação $\theta(d) \in \{Rng(d,\xi), k\text{-}NN(d,k)\}$ para indicar indistintamente qualquer dos dois operadores de comparação por similaridade usando o descritor d, e também para distingui-lo de um operador de comparação θ genérico, aplicável a qualquer domínio $\mathbb{A}$.

A recuperação dos dados armazenados é feita pelos **operadores de busca**

. Existem diversos operadores de busca comumente utilizados em consultas, tais como os operadores de seleção $\sigma_{(c)}$, junção $\overset{(c)}{\bowtie}$ e projeção $\pi_{\{l\}}$. A seleção é um dos operadores de busca que permitem expressar explicitamente as comparações (c) entre os valores dos atributos armazenados nas diversas relações da base de dados usando **condições de consulta** sobre os atributos. Uma condição de seleção c é expressa como $(\mathsf{A}\ \theta\ a_j)$ e indica a comparação do valor a_i do atributo A em cada tupla t_i da relação com um valor a_j.

Com a emergência do suporte a dados multimídia em SGBDs, os operadores de busca baseados em operadores de comparação por similaridade vêm despertando interesse crescente, principalmente para a recuperação por conteúdo (SILVA; AREF; ALI, 2010; BUDIKOVA; BATKO; ZEZULA, 2012; TRAINA *et al.*, 2010; SILVA *et al.*, 2010; BARIONI *et al.*, 2010; LU *et al.*, 2017; JR. *et al.*, 2019). Um operador de busca por similaridade deve recuperar os elementos que atendem a um determinado critério expresso por um operador de comparação por similaridade. Dessa maneira, quando um operador de seleção ou junção é executado usando um operador de comparação por similaridade, o resultado é uma busca por similaridade.

Um **operador de seleção** seleciona as tuplas de uma relação que atendem a uma condição

de consulta. Ele é expresso como $\sigma_{(\mathsf{A}\ \theta\ a_q)}\mathsf{T}$, onde T é uma relação, A é um atributo dessa relação, θ é um operador de comparação válido no domínio de A e a_q é um elemento do mesmo domínio do atributo A. A condição de seleção $(\mathsf{A}\ \theta\ a_q)$ indica a comparação do valor $a_i = t_i[\mathsf{A}]$ do atributo A em cada tupla t_i da relação com o valor a_q. A operação de seleção recupera todas as tuplas da relação T em que o valor armazenado no atributo A resulta num valor `Verdade` para a condição (resultados `Falso` e `Desconhecido` descartam a tupla).

Um **operador de seleção por similaridade** $\sigma_{(\mathsf{S}\ \theta(d)\ s_q)}\mathsf{T}$ compara atributos complexos S usando uma condição de consulta por similaridade. Nesse caso, o valor de comparação s_q dado na consulta é chamado "**centro da consulta**". A operação de seleção por similaridade recupera, dentre as tuplas da relação, aquelas cujos valores armazenados no atributo S atendam à condição de consulta em relação ao centro da consulta. O operador de comparação $\theta(d)$ é o comparador por abrangência ou pelos k-vizinhos mais próximos: $\theta(d) \in \{Rng(d,\xi), k\text{-}NN(d,k)\}$.

Uma **consulta por abrangência**, expressa como $\sigma_{(\mathsf{S}\ Rng(d,\ \xi)\ s_q)}\mathsf{T}$, recupera as tuplas t_i de T tal que o valor $s_i = t_i[\mathsf{S}]$ do atributo S está a até a distância máxima ξ do valor complexo s_q medida pelo descritor d. Uma **consulta por vizinhança**, expressa como $\sigma_{(\mathsf{S}\ k\text{-}NN(d,\ k)\ s_q)}\mathsf{T}$, recupera as tuplas de T tal que o valor s_i do atributo S é um dos k valores mais próximos a s_q no espaço de busca formado pelo descritor d sobre os valores armazenados no atributo complexo S.

2.2 Padronização de Dados

Encontrar correlações e padrões em dados é uma tarefa realizada em diversas áreas. Buscam-se correlações na economia, no marketing e claro, na medicina. O interesse por se encontrar uma eventual droga, procedimento, exame ou qualquer variável que possa auxiliar no tratamento de pacientes ou no desenvolvimento de novos medicamentos e tratamentos é absolutamente desejável e, não raramente, tem potencial para contribuir com projetos de diferentes partes do mundo. Contudo, para que se possa correlacionar tais dados, é necessário que os mesmos estejam armazenados seguindo algum tipo de padronização. Existem milhares de drogas, exames, doenças etc. que podem ser armazenados em um banco de dados de forma totalmente diferente quando se compara dois ou mais hospitais, por exemplo.

Aqui, no Brasil, para ilustrar, um padrão bastante conhecido para indicar achados clínicos é o Código Internacional de Doenças em sua versão 10, o CID-10. Com ele, todas as doenças estão categorizadas de maneira padronizada para que elas sempre tenham um mesmo código identificador, padronizando e classificando todos os pacientes que delas sofrem, independentemente de qual é a instituição o ou quem é o profissional responsável pelo diagnóstico. Essa tendência de padronização em nível nacional ocorre no Brasil devido ao padrão ter sido instituído pelo Sistema Único de Saúde - SUS. Contudo, em uma esfera mundial, nem todos os países seguem um mesmo padrão e, com isso, nem sempre é viável executar uma análise utilizando o padrão CID-10 integrando dados de diversas instituições, pois possivelmente haveriam dados de

instituições cujos dados são armazenados segundo outra nomenclatura.

Não apenas os nomes de doenças devem ser padronizados. A utilização de drogas, por exemplo, segue uma terminologia que em geral, varia enormemente quando se comparam diversas unidades de saúde, o que dificulta a utilização de dados de diferentes hospitais em um mesmo estudo ou ainda, que uma análise seja aplicada em outra base de dados. Além disso, a própria estrutura das informações mantidas em cada instituição geralmente atende à modelagens particulares a cada uma, de maneira que a integração de dados de instituições independentes em geral requer um tratamento separado.

2.2.1 Modelos Conceituais

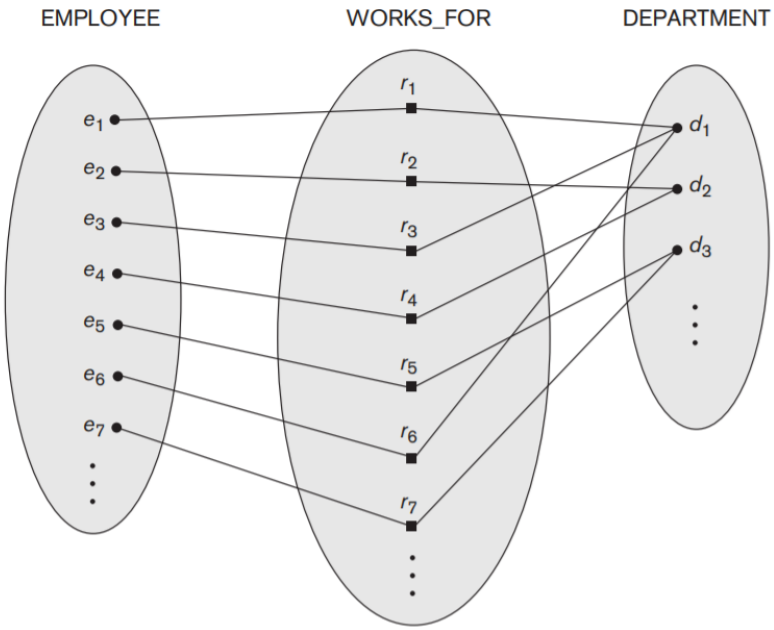
Existem diferentes formas de se organizar um conjunto de dados a fim de se obter uma estrutura abstrata que, quando implementada, permite a escrita e a recuperação de informação sejam feitas de forma eficiente e confiável. (ELMASRI; NAVATHE, 2010) descreve um modelo de dados como sendo uma coleção de conceitos que podem ser usados para descrever a estrutura de uma base de dados e, por estrutura de uma base de dados, refere-se aos tipos, relações e variáveis que são aplicados aos dados. A grande maioria dos modelos de dados também prevê um conjunto de operações básicas para especificar a recuperação de dados e atualização dessa base. Os **modelos de dados** são essencialmente um conjunto de regras que, quando aplicadas à análise de um empreendimento ou a um conjunto de informações, permitem criar a **modelagem** desse empreendimento/informações.

Dentre os diversos modelos conceituais que foram desenvolvidos ao longo da história da Computação, um dos mais tradicionais e populares é o modelo Entidade-Relacionamento (ER) 1. Uma **E**ntidade nada mais é que um objeto, um conceito, algo que se está representando através de seus dados, como por exemplo, um paciente em particular, um determinado hospital, ou mesmo um objeto com existência apenas conceitual, como por exemplo uma consulta. Cada entidade possui e é descrita pelos seus atributos, ou seja, as características que a definem. No nosso exemplo, uma entidade paciente pode ser representada por seus atributos como idade, gênero, o endereço residencial e nome. Por sua vez, cada atributo possui um determinado valor. De forma prática e ilustrativa, um conjunto de entidades pode ser implementado como uma tabela. Já os **R**elacionamentos podem ser compreendidos informalmente como sendo uma associação entre entidades, cada qual também podendo ter atributos. Por exemplo, uma entidade Paciente pode ter um relacionamento Recebe com uma entidade 'Procedimento Médico', e que tem por exemplo como atributos a data e o custo do procedimento.

Para se construir uma modelagem ER, as regras que regem tal modelo devem ser, obviamente, aplicadas. Isso resulta em uma série de características que podem ou não ser vantajosas, a depender exclusivamente dos recursos disponíveis, sejam eles físicos – como espaço de armazenamento – ou não – como a complexidade para recuperar informação ou o tempo necessário para fazê-lo. O Modelo ER foi concebido e tem como sua principal vantagem o

fato de que ele gera modelagens onde os dados podem ser mantidos normalizados, o que permite controlar a consistência dos dados representados.

Figura 1 – Modelo Entidade Relacionamento (ER).



Fonte: (NAVATHE, 2017)

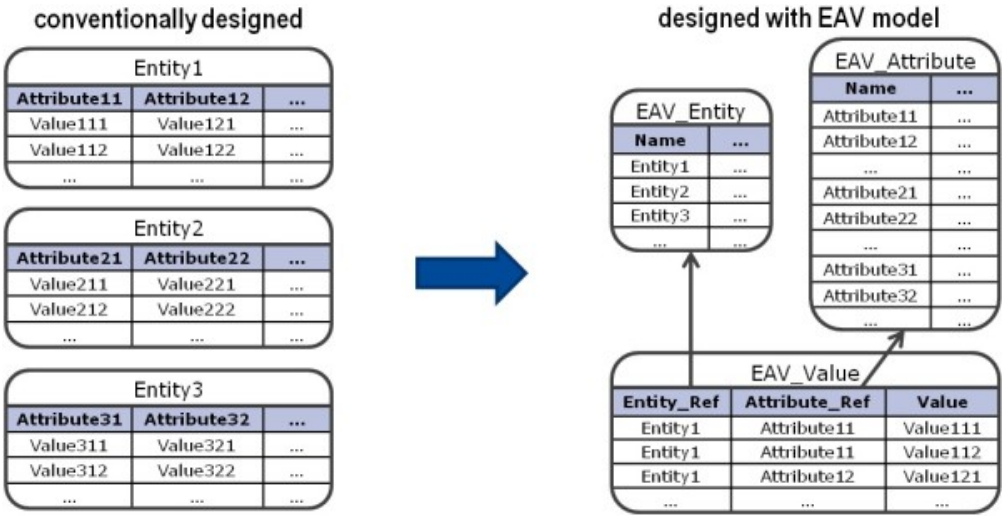
2.2.2 Modelo EAV

MEAV (Modelo Entidade-Atributo-Valor) é um modelo de dados ou representação que, dentre diversas características que serão apresentadas a seguir, preza, especialmente, a flexibilidade das modelagens resultantes. Ele é amplamente utilizado em áreas onde existe grande necessidade de adaptação e evolução da estrutura dos dados armazenados, ou onde há ausência de uma estrutura centralizada de padronização ou previsibilidade sobre qual tipo de dado será necessário armazenar. Esse modelo tende a se encaixar muito bem aos contextos médicos, especialmente para a representação das informações relacionadas a um paciente que é tratado em diversas instituições diferentes ao longo do tempo. Esses modelos podem ser comparados a matrizes esparsas, fazendo-se uma analogia com tal estrutura matemática.

De acordo com (NADKARNI *et al.*, 1999), a representação MEAV é, primeiramente, uma forma de simplificação de um esquema físico de um banco de dados, a ser usado quando a generalização de algumas tabelas possa ser benéfica. Pode-se descrever a forma usual de um modelo EAV como uma tabela com três colunas – uma coluna para a **E**ntidade (objeto a ser descrito) contendo uma identificação (ID), um para **A**tributo/parâmetro ou um ID de atributo que possivelmente aponta para uma tabela com as descrições dos atributos) e, finalmente, uma tabela que contem os **V**alores desses atributos. Nessas estruturas, ao contrário dos modelos ER, os dados são descritos de forma explícita, no que podemos considerar tabelas de metadados – ou

seja, cujo conteúdo não é o dado em si, mas informações sobre o que aquele dado é. A figura 2 ilustra um exemplo de modelo EAV.

Figura 2 – Exemplo de Modelo EAV



Fonte: Löper *et al.* (2013).

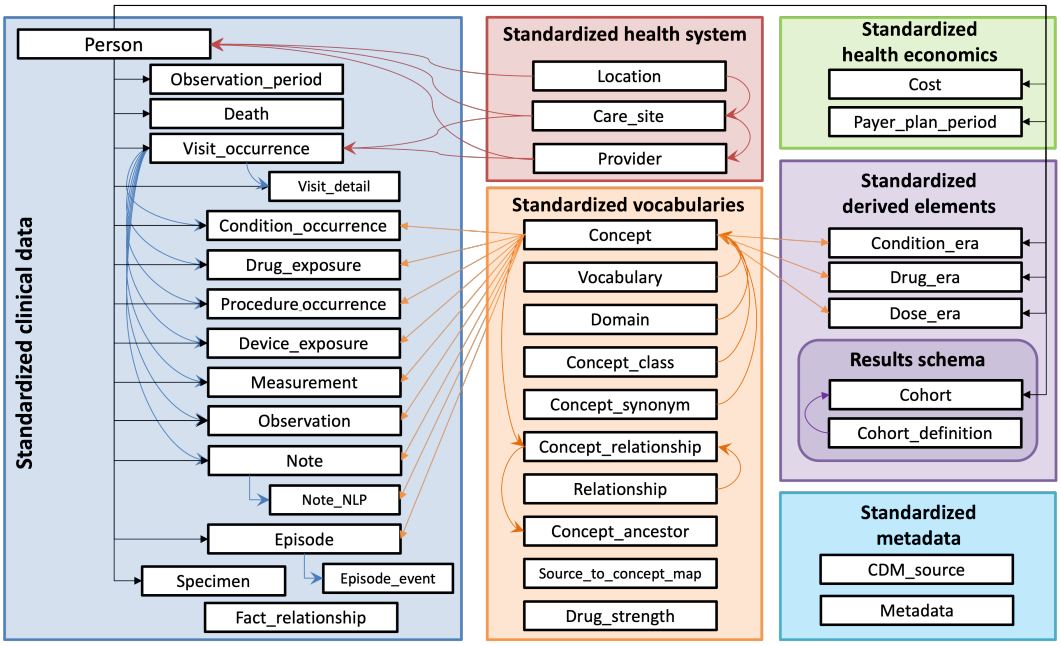
Modelagens baseadas no modelo EAV são reconhecidamente mais complexas e difíceis de se trabalhar, exatamente devido às suas características de falta de padronização e que, embora tenham os benefícios já citados, tornam a recuperação de informação mais trabalhosa do que em modelos ER. Uma das desvantagens de usar esse modelo é notada quando uma consulta precisa ser executada sobre valores de diversos atributos. Por exemplo, realizar uma consulta que retorne os pacientes com idade acima de 45 anos que realizaram um Eletrocardiograma e que também apresentaram níveis elevados de leucócitos exigirá uma escrita de consulta bastante complexa e que será menos eficiente (considerando os tempo de execução da consulta) que em uma estrutura ER. Isso ocorre porque num modelo ER, os valores de todos os atributos de um mesmo paciente estará em uma mesma tupla, ao passo que em um modelo EAV cada valor estará em uma tupla diferente, tendo como único elemento de ligação o valor do identificador do paciente. De fato, o incremento no tempo de consulta em modelagens EAV é tão relevante que diversos trabalhos na literatura se dedicam a essa mensuração. Em (CHEN *et al.*, 2000), as consultas a bases de dados foram de 3 a 5 vezes menos eficiente (considerando tempo de execução das consultas) em modelagens EAV quando comparadas a modelagens ER equivalentes. No entanto, o modelo EAV permite representar numa mesma tabela dados de um mesmo paciente provenientes de diferentes instituições de saúde, que adotem representações diversas para os dados.

Com o uso em larga escala de modelos EAV no âmbito da medicina, diversas abordagens foram propostas para tornar a tarefa de recuperação de informação mais eficiente do ponto de vista do usuário, seja ele um DBA ou um especialista de outra área, como um médico, por exemplo, sendo uma delas, a utilização de Técnicas de Visualização de Dados, que serão discutidas no próximo capítulo.

2.2.3 OMOP-CDM

Bases de dados que seguem o padrão de modelagem OMOP (CDM5-4doc, 2022) (The Observational Medical Outcomes Partnership) visam a **interoperabilidade** dos dados entre diversas instituições de saúde, e são construídas seguindo uma modelagem híbrida, combinando o Modelo Entidade-Relacionamento (ME-R) e o Modelo Entidades-Atributo-Valor (MEAV). Em tal modelagem, a estrutura fundamental dos dados comum a todas as instituições é modelada seguindo o ME-R, de maneira que ela sempre segue o mesmo padrão, independentemente das particularidades da instituição "dona dos dados". As entidade dos Conjuntos de entidade modelados podem ser representadas com atributos que se caracterizam como objetos complexos. Já a estrutura dependente de cada instituição é modelada seguindo o MEAV, e assim os atributos que são específicos de cada instituição/sistema são acrescentados como novas tuplas, associando às respectivas entidades os atributos que aquela instituição/sistema trabalha, com os respectivos seus valores medidos. A figura 3 mostra o esquema geral do Modelo Comum de Dados em sua versão 5.4, onde os retângulos brancos representam os conjuntos de entidades contemplados.

Figura 3 – Esquema OMOP-CDM v5.4



Fonte: cdm54 (2022).

A padronização OMOP-CDM é administrada pela iniciativa ODHSI, uma organização para a colaboração e interoperabilidade de dados médicos que possui uma comunidade espalhada por mais de 19 países e dados de mais de meio bilhão de pacientes até a presente data. O Brasil está iniciando um processo de integração do padrão do SUS para o Snomed, baseado no OMOP-CDM, capitaneado pelo hospital InCor e financiado pelo projeto temático do qual esta proposta de mestrado faz parte. A proposta da iniciativa ODHSI é criar e manter um padrão de conversão para rotulagem de dados médicos, ou seja, criar uma padronização de nomenclatura

de dados que podem até mesmo ser interoperáveis com outros padrões, como o CID-10 ou o *Systematized Nomenclature of Medicine* (SNOMED), por exemplo. De fato, a iniciativa ODHSI integra no OMP-CDM esses diversos padrões sob uma arquitetura unificada.

A modelagem de dados médicos de fontes originais (oriundas do hospital ou qualquer base médica primária) para o OMOP-CDM exige que um amplo trabalho de avaliação, limpeza, validação e transformação (PUTTMANN *et al.*, 2022) e também como apresentado em (LIMA *et al.*, 2019), que descreve as etapas de modelagem da base utilizada neste trabalho.

2.2.4 Conclusões

Neste capítulo foram apresentados os conceitos necessários para se descrever e comparar objetos complexos. Para isso, é necessário, primeiramente, que se descreva os objetos utilizando uma associação de atributos - como por exemplo, caracterizar um paciente através de seus dados demográficos e de seus exames médicos. Uma vez que sejam descritos, define-se como esses objetos serão comparados. Para isso, foram apresentados os conceitos de similaridade e espaços métricos, bem como os operadores de comparação, busca e seleção. Finalmente, considerando tratar-se de um banco de dados, faz-se necessária a apresentação dos conceitos que formam a base do modelo relacional e particularidades da base que armazena os dados do InCor, instituição que detém os dados que se pretende tratar neste projeto de mestrado.

No capítulo seguinte, serão apresentados os conceitos relacionados à Visualização de Dados e como se pretende que ela auxilie o processo de compreensão dos dados.

CAPÍTULO

3

VISUALIZAÇÃO DE DADOS

A Visualização de Dados se relaciona com diversas áreas do conhecimento, indo desde a matemática e computação, à neuro-ciência, psicologia e artes e até os limites da própria capacidade humana de percepção do conhecimento. Diante dessa característica multidisciplinar do tema, iremos nos ater ao estudo científico da visualização de dados e informação, muito embora a avaliação de um usuário humano esteja prevista como chancela de sucesso ao final deste trabalho. De acordo com (WILLIAMS; SOCHATS; MORSE, 1995), "visualização é um processo cognitivo realizado por humanos através da formação de uma imagem mental de um domínio do espaço". Em Computação e Ciência da Informação, a visualização, de forma mais específica, pode ser descrita como a representação visual de um domínio de espaço utilizando gráficos, imagens, sequências animadas e sons para apresentar dados, estruturas e comportamento dinâmico de grande e complexos conjuntos de dados, que por sua vez representam sistemas, eventos, processos objetos e conceitos. - Tradução livre da autora.

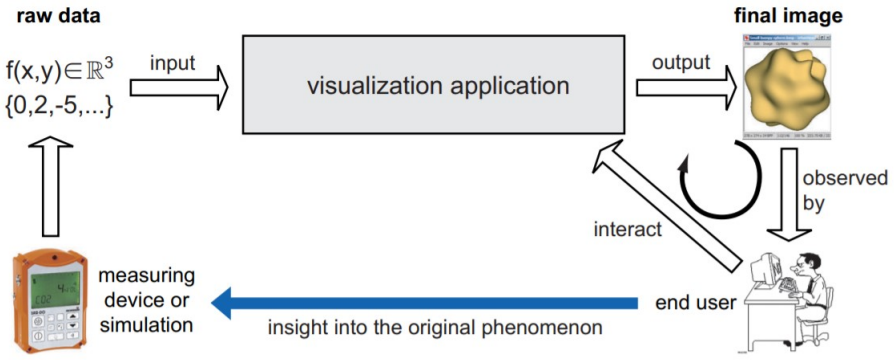
Um vocábulo importante bastante atrelado ao campo de visualização é o *insight*, um “lampejo” para algo que não havia sido notado através de outros métodos de análise dos dados. Um *insight* pode também ser descrito como o processo de expansão do conhecimento sobre determinado conjunto de dados, através de artifícios gráficos tais como gráficos, grafos, variação de cores e formas. Prover o *insight* para o usuário ou analista de dados é extremamente importante, sendo, muitas vezes, a grande meta do trabalho de visualização. Além disso, esse processo na maioria das vezes é feito de forma cíclica. Os dados que são coletados e se tornam o corpo da visualização geram conhecimentos que podem ser reincorporados como dados na base original, como mostrado na figura 4.

Ainda sobre a descoberta de conhecimento, de forma ilustrativa, podemos formular questões sobre um conjunto de dados que, quando respondidas, nos conduzem aos já citados

insights. A figura 5 divide e organiza essas questões, que se subdividem em 2 tipos. O primeiro deles são questões concretas (perguntas específicas) sobre determinado fenômeno, processo ou conjunto de dados e que em geral podem ser respondidas submetendo-se os dados a um processo de cálculo de medidas estatísticas. O propósito da visualização, neste contexto é responder de maneira intuitiva à questões como:

- Qual é o valor mínimo, máximo ou os *outliers* de determinado conjunto de dados e para quais dados estes valores ocorrem?
- Qual é a distribuição dos valores de determinado conjunto de dados?
- Os valores de diferentes conjuntos de dados mostram alguma correlação?
- Quão bem os valores de determinado conjunto de dados se encaixam em um dado modelo matemático ou padrão?

Figura 4 – Ciclo de *Insight*

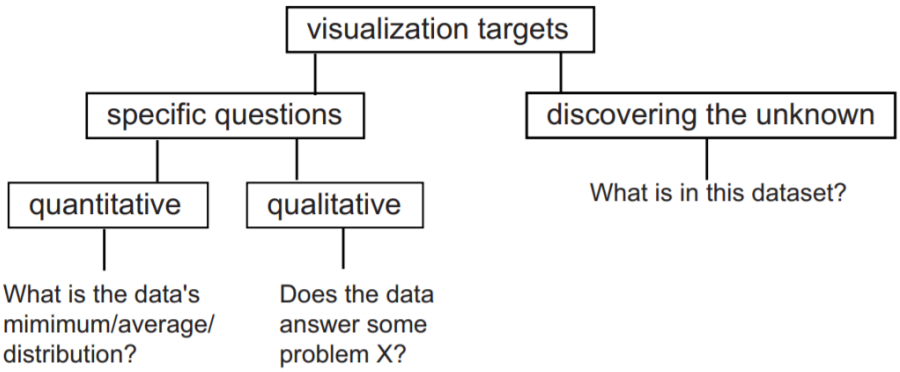


Fonte: (TELEA, 2015)

Apesar de temperatura ser um número simples e não precisar de auxílio visual para que seja claramente compreendido, uma sequência de temperaturas representada por esses números será quase sempre mais rapidamente compreendida se representada graficamente.

Já o segundo tipo de pergunta (referente à descoberta de conhecimento) corresponde justamente ao grupo de perguntas que não sabemos que temos que fazer. São aquelas que se referem ao conhecimento que não sabemos que temos, ou que podemos extrair de determinado conjunto de dados. Um pesquisador ou analista pode começar sua busca com as perguntas do primeiro tipo mas, em um momento posterior, empregando diferentes tipos de visualização, poderá se deparar com associações entre os dados que lhe darão mais clareza de como aquele conjunto se comporta.

Figura 5 – Tipos de objetivos.



Fonte: Telea (2015).

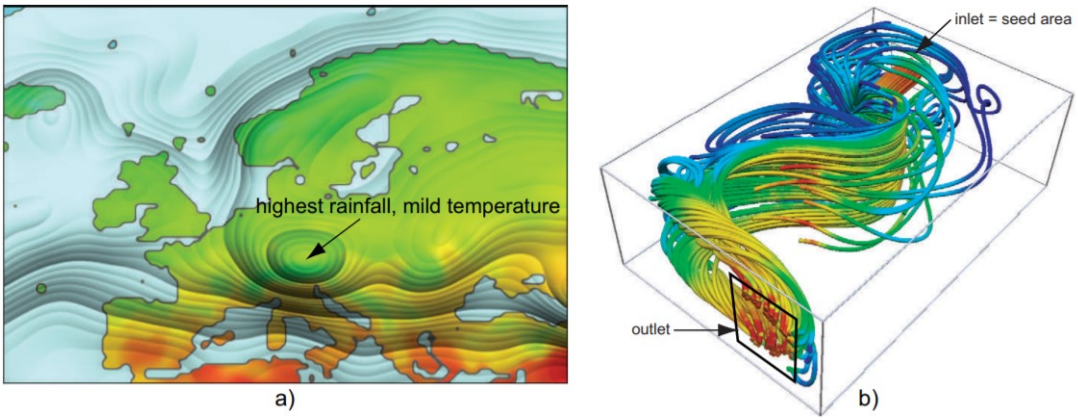
3.1 Subdivisões da Visualização de Dados

A Visualização Científica (comumente abreviada como scivis) – atualmente também conhecida por visualização de dados espacial, despontou como um tema independente no final dos anos 80, em decorrência do aumento da quantidade de dados gerados pela execução de simulações numéricas computacionais de processos físicos, tais como escoamento, convecção de calor e deformação de materiais. A Visualização Científica pode ser compreendida como um conjunto de técnicas para representação visual de dados de qualquer natureza que tenham uma distribuição geométrica intrínseca. De fato, a palavra "Científica" que compõe o nome da área, deriva do objetivo inicial de representar as simulações científicas, que em sua maioria, eram modelos tridimensionais. Alguns exemplos de aplicação desse tipo de visualização podem ser observados nas áreas de Arquitetura, Meteorologia, Medicina e Biologia, onde a ênfase é em elementos reais como volumes, superfícies, iluminação e outros, inclusive com algum tipo de evolução temporal.

Conforme ilustrado e exemplificado no livro (TELEA, 2015), diversos exemplos de visualização científica podem ser dados. A figura 6 exibe a visualização da temperatura média e chuva sobre a Europa no mês de Julho durante o período de 1960 a 1990. Já o item b da figura 6 é capaz de ilustrar o escoamento de um fluido em uma caixa. Os tubos são coloridos com relação à velocidade de escoamento (azul para lentos, vermelho para rápidos).

Além de dados espaciais - que estão tratados pela Visualização Científica, muitos outros tipos de dados como árvores, grafos, e redes necessitam também de alguma forma de visualização. Apesar desses tipos de dados também necessitarem de alguma representação espacial para serem desenhadas em um pedaço de papel, por exemplo, as informações espaciais são atribuídas aos elementos de dados durante a construção da visualização ao invés de serem fornecidas pelos próprios elementos de dados. Esses tipos de dados, incluindo também tabelas,

Figura 6 – Visualização Científica

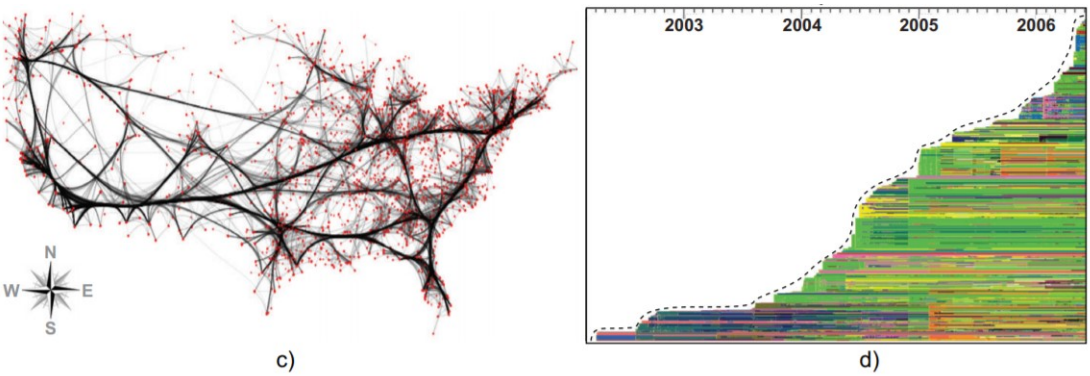


Exemplos de imagens de Visualização Científica (a) Mapa de distribuição de precipitação sobre a Europa no mês de Julho entre 1960 e 1990; (b) Fluxo de escoamento de um fluído em uma caixa.

Fonte: Telea (2015).

séries temporais, documentos e código-fonte do software, formam o objeto de estudo de uma área separada, chamada Visualização de Informações ou Informações Resumidas, que pode ser definida como um conjunto de técnicas para representação visual de dados de qualquer natureza que não necessariamente depende de uma organização geométrica intrínseca. O aumento no número, tamanho e tipos de artefatos digitais causados pela sociedade da informação na última década, geralmente chamados "*big data*", tem sido um catalisador particularmente importante do crescimento do interesse nesta subárea, a visualização de informações (também encontrado comumente na literatura como Infovis, conforme exibido na figura 7).

Figura 7 – Exemplos de Infoviz



Fonte: Telea (2015).

Por último, a Visualização Analítica que surge como uma ponte entre as duas primeiras subáreas. De forma sucinta, essa subárea híbrida tem como objetivo fornecer técnicas e ferramentas que suportem seus usuários finais que são, em geral, especialistas em determinada área do conhecimento em seu raciocínio analítico por meio de interfaces visuais interativas. (THOMAS,

2004) A visualização analítica não é comparável com scivis e infovis: ela integra visualização de dados (que num dendograma de classificação de áreas estaria um nível acima de ambas) com mineração de dados: A visualização Analítica tem o objetivo de usar técnicas de visualização de dados para descobrir informações nos dados.

Apesar de haver essa distinção com fronteiras não muito claras entre a Visualização Analítica e a scivis e infovis, existem aspectos que as tornam diferençáveis. Por exemplo, pode-se dizer que a Visualização Analítica se concentra em todo o processo de "criação de sentido", que começa com a aquisição dos dados e continua através de vários cenários repetidos e refinados – já que essa subárea tem como objetivo, principalmente, garantir que os usuários explorem diferentes pontos de vista ou hipótese e termina apresentando o *insight* adquirido pelo usuário final sobre os fenômenos de interesse. Também pode-se dizer que a Visualização Analítica é caracterizada por uma combinação de tecnologias e ferramentas de análise, mineração e visualização de dados. Assim, a mineração de dados, várias visualizações e a inspeção visual interativa e iterativa dos dados são componentes inseparáveis da análise visual.

3.2 Visualização de Dados Médicos

Dadas as especificidades de cada base de dados e os objetivos desejados em cada processo de visualização, em geral é necessário que se produzam soluções desenvolvidas especificamente para cada situação (inclusive, comerciais) - isto é, de propósito específico (*tailor made*). Por exemplo, a visualização de informações a respeito de pacientes que são submetidos a uma ressonância magnética diferem das visualizações indicadas para pacientes que realizam exames de colesterol ou Eletrocardiogramas e, mais ainda, diferem entre si as motivações e as técnicas para se desenvolva tal visualização. Desta forma, diversas técnicas são aplicadas à analise de dados médicos, não havendo, tanto quanto é de nosso conhecimento, padronização, recomendação nem predileção por determinado tipo.

Diversos exemplos de aplicação de Técnicas de Visualização de Dados médicos podem ser encontradas na literatura, mas em geral são aplicadas na visualização de casos específicos (AZEVEDO-MARQUES *et al.*, 2017; HAAK; PAGE; DESERNO, 2016). Por outro lado, a visualizações de dados médicos, tal como é o interesse deste projeto de mestrado, tendo a ser específica para aplicações pré-determinadas, e assim são poucos os trabalhos genéricos que avaliam a aplicabilidade de técnicas de visualização para dados médicos em geral. Por exemplo, no trabalho desenvolvido por (KHALID *et al.*, 2014), pode-se comparar a efetividade do uso de técnicas de visualização de Coordenadas de Estrela para auxiliar na identificação de correlação de agrupamentos entre determinados atributos.

3.3 O Software de Visualização `Tableau` e a linguagem de consulta para visualização

O software comercial de visualização de Dados `Tableau` versão *Desktop* (TABLEAU, 2022) foi a ferramenta utilizada para a geração dos gráficos deste trabalho. A escolha deste programa foi pautada pelo amplo uso feito pela comunidade de Análise de Dados, tanto acadêmica quanto de mercado, pelo nível de excelência nas avaliações de institutos de referência, como o Gartner (GARTNER, 2022) além da vantagem de haver uma parceria com esta instituição de ensino, que libera uma licença gratuita para estudantes.

O Tableau é um software cujo objetivo é também muito semelhante ao que propõe este trabalho - oferecer ao usuário uma forma de analisar os dados sem que este precise dominar uma linguagem de consulta ou se preocupar com a arquitetura em que este dado está sendo armazenado. Evidentemente, o conhecimento e o domínio de uma linguagem tem um valor inestimável para o desenvolvimento da ciência e para questões mais direcionada à recuperar a informação em forma mais bruta mas, oferecer uma forma democrática de descoberta do conhecimento a especialistas de outro domínio é imperativo para o progresso das demais frentes do saber, como medicina, foco deste mestrado. Com este mesmo pensamento, o `Tableau` foi projetado tendo como paradigma a utilização de uma interface visual que, quando manipulada, gera automaticamente consultas SQL para o software, traduzindo as necessidades do utilizador sobre as bases de dados.

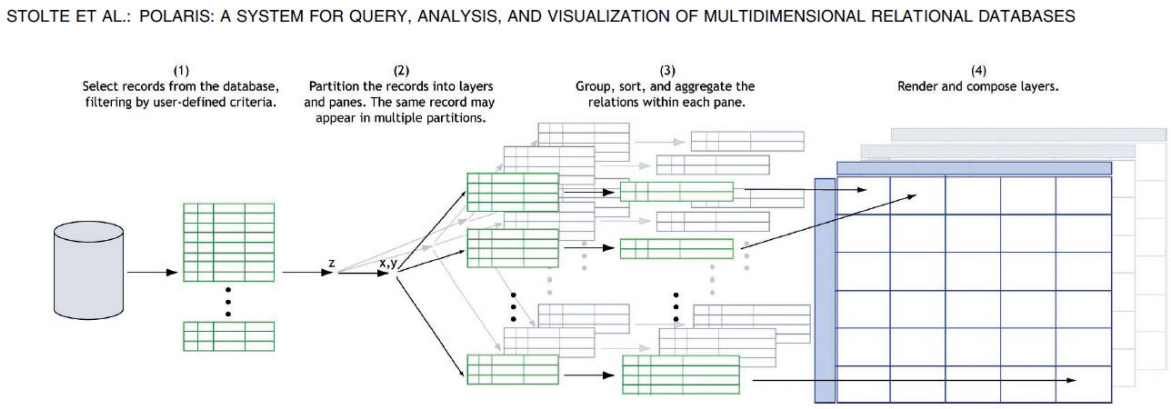
A essa forma de traduzir as necessidades visuais do usuário foi dada inicialmente o nome de Formalismo de Polaris (STOLTE; TANG; HANRAHAN, 2002), para mapear onde os dados necessários para a consulta estão e, depois, em construir uma consulta que responda a esses critérios, que são chamados de filtros. O Paradigma de Polaris evoluiu para o que se tornou a linguagem declarativa VizQL™ (Visual Query Language for Databases) (HANRAHAN, 2006). O VizQL™ usa conceitos de SQL e MDX (para manipulação de cubos OLAP) para criar elementos visuais iterativos que serão explorados neste trabalho nas próximas seções.

3.3.1 Carregando a fonte via conexão com Banco de Dados

A forma mais direta de se utilizar os dados de um banco no Tableau é utilizando uma conexão direta entre a base e a plataforma de visualização. O sistema oferece uma vasta lista de conectores, incluindo ferramentas de gerenciamento de *datawarehouses* de diferentes empresas como `Google`, `Oracle` e `Amazon`. Conexões utilizando os padrões ODBC e JDBC também são possíveis.

É altamente recomendado utilizar uma conexão direta sempre que for necessário analisar dados que sofrem alterações com frequência (tipicamente em um modo de operação transacional *OnLine Transactional Processing – OLTP*), e quando as tabelas necessárias para a visualização já estiverem em uma estrutura viável. Ressalta-se que, apesar disso, é possível, a partir do `Tableau`,

Figura 8 – Mapeamento de dados no paradigma polaris.



Fonte: Stolte, Tang e Hanrahan (2002).

realizar a modelagem de novas métricas e dimensões, e especialmente quando se utiliza diversas colunas, conforme será apresentado nos capítulos seguintes.

3.3.2 Carregando dados via exportação

Uma outra abordagem para a inserção de dados consiste na importação dos dados a partir da base de dados externa para dentro do Tableau. Nesta modalidade, a plataforma aceita os formatos mais usuais como JSON, CSV (arquivo de texto com valores separados por vírgula), PDF e Excel. Essa foi a abordagem utilizada neste trabalho, já que desejávamos realizar uma análise em um recorte estático de dados, não sendo necessário manter uma conexão entre o banco e o `Tableau`.

Uma vez que os dados estejam carregados (neste exemplo usando um arquivo `.csv` gerado a partir da consulta base deste trabalho que será explorada com mais detalhes nas próximas sessões, é possível realizar o *parsing* das colunas bem como explicitar a tipagem das colunas, indicando se os valores são, por exemplo, números, texto, moeda, data, etc. Além disso, é possível alterar o nome das colunas e criar conexões entre tabelas (*joins*), como mostrado na figura 9.

3.4 Modelando métricas e dimensões no `Tableau`

Métricas e dimensões são nomenclaturas utilizadas amplamente na área de Inteligência de Negócios (BI) para caracterizar atributos ou medidas quantitativas nos dados. Métricas podem ser definidas como medidas quantitativas que indicam alguma propriedade dos elementos de dados e representada como atributos do conjunto de dados os quais, por sua vez, é uma dimensão deste conjunto de dados. Seguindo o exemplo-mote deste capítulo, a condição do paciente (infarto) descrita pela coorte que escolhemos e apresentamos no item 4.2.2 é uma dimensão e a

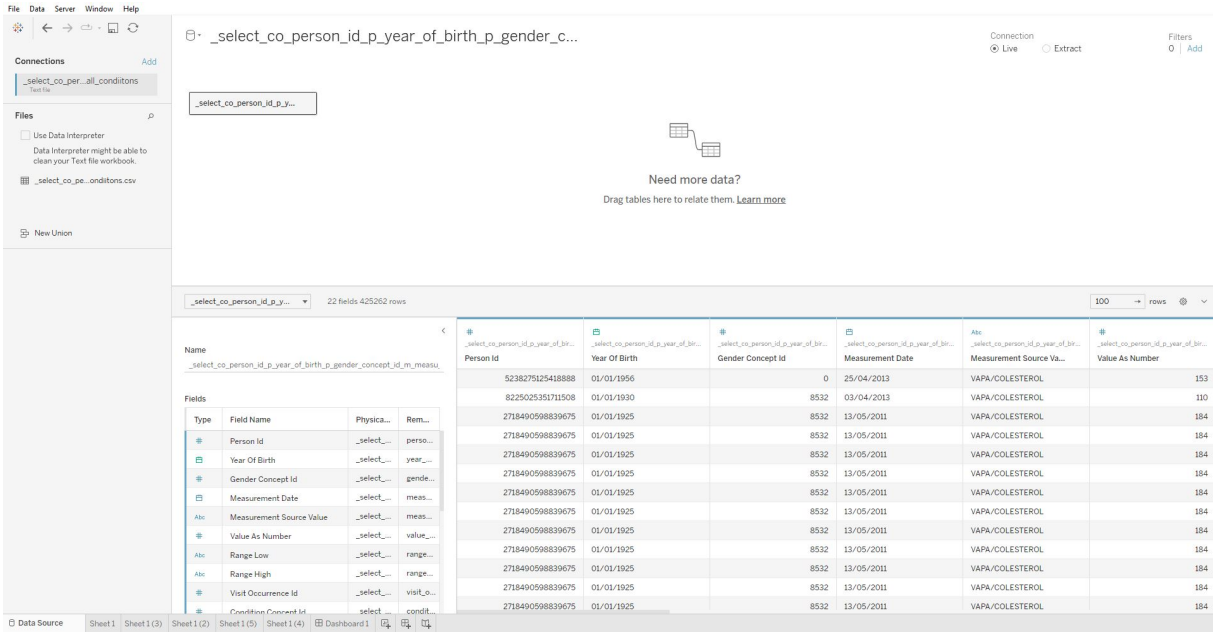


Figura 9 – Interface para adição de Fonte de Dados no `Tableau` v 2022.2

Fonte: Elaborada pelo autor.

média de exames (de colesterol) que esse paciente realiza por ano é uma métrica.

As métricas e dimensões derivadas, isto é, que utilizam métricas e dimensões primárias da fonte de dados podem e devem ser criadas dentro da plataforma de visualização, para que se possa obter uma maior compreensão do comportamento dos dados. Não há, contudo, uma definição pré-estabelecida da quantidade ou de quais métricas devem ser criadas. Cabe ao analista de dados ou responsável pelo estudo compreender as diversas facetas do problema e determinar quais modelagens podem ser úteis. Temos por objetivo aqui, portanto, guiar o utilizador sobre as capacidades do software e de ilustrar a metodologia utilizada para a exploração de um conjunto de dados em particular, conforme será feito no próximo capítulo.

No `Tableau`, a forma mais comum para se criar novas métricas e dimensões é através dos campos calculados. Isso permite que o usuário crie uma nova coluna no seu conjunto de dados, aplicando sobre ele regras, cálculos matemáticos, filtros e conversões de tipos.

Os campos calculados podem ser computados em nível de tabela, de linha ou da visualização. Alguns recursos importantes que o leitor deve ter conhecimento para explorar o potencial da plataforma são as expressões em *Level Of Detail* (conhecidas como LOD) e as *Table Calculations*. As LOD são expressões que permitem usar dados em uma granularidade diferente daquela usada na visualização - nem sempre desejamos utilizar o nível de detalhes mais granular na visualização mas podemos querer entender detalhes mais granulares de um bloco da visualização como, por exemplo, em um gráfico cujo visual esta agrupado por tipo de doenças, termos discriminados a divisão dessas doenças por sua ocorrência geográfica. Ora, poderíamos incluir a dimensão geográfica na visualização mas nem sempre mudar a visualização é o que

desejamos para compreender um comportamento mais detalhado dos dados que estão agrupados.

Já as *Table Calculations* são cálculos que devem ser executados sobre o resultado da consulta que gerou a visualização - ou seja, estão na mesma granularidade que os dados do visual. Um exemplo ilustrativo de *Table Calculations* é incluir na visualização os totais por linhas de uma tabela.

Apesar de haver uma ampla literatura disponível sobre LODs e *Table Calculations*, o objetivo desta seção é familiarizar o leitor com os recursos mais frequentes e necessários no uso do Tableau, bem como apresentar uma revisão bibliográfica dos conceitos sob os quais o software utilizado se sustenta. Mais detalhes e aplicações deste recursos serão abordados nos próximos capítulos.

3.5 Criando Gráficos e outros elementos visuais

A visualização de dados pode ser feita através do uso de diferentes gráficos ou estruturas, conforme discutido no capítulo 5. A escolha do melhor tipo de visualização, juntamente com a determinação dos atributos, dimensões e métricas forma o núcleo fundamental de elementos de exploração de dados baseada em objetos visuais e, portanto, é imprescindível que uma escolha criteriosa seja feita para que se conquiste os objetivos fundamentais da visualização: A inteligibilidade das informações, a capacidade de destacar padrões, tendencias ou correlações ou determinado sub-conjunto de elementos que atendem à um critério ou a uma hipótese que deve ser previamente formulada pelo pesquisador ou usuário.

3.5.1 Os tipos de gráficos do Tableau

O `Tableau` dispõe de diversos tipos de gráficos padrão. Entre eles estão: Linha, Barra, Pizza, Mapas, Mapa de Densidade, Scatter Plots, Gráficos de Gantt e de Bolhas, Histograma, Gráfico em Bullet, Tabela de Destaque, Treemap, Box/Whisker, e Gráfico de Vela. Alguns desses tipos podem ser vistos como exemplo para apreciação do potencial visual da ferramenta nas figuras 10 e 11.

3.5.2 Adicionando guias analíticas aos gráficos

Além do tipo mais adequado de gráfico, o uso de guias analíticas podem ajudar grandemente na compreensão de um dado quando comparado ao conjunto. Por exemplo, o uso de linhas visuais indicando os limites dos valores médios para uma determinada distribuição de valores pode, rapidamente, classificar uma medida como sendo ou não pertencente a um grupo de interesse (por exemplo, que determinado paciente tem um valor para um exame que está acima do valor médio esperado).

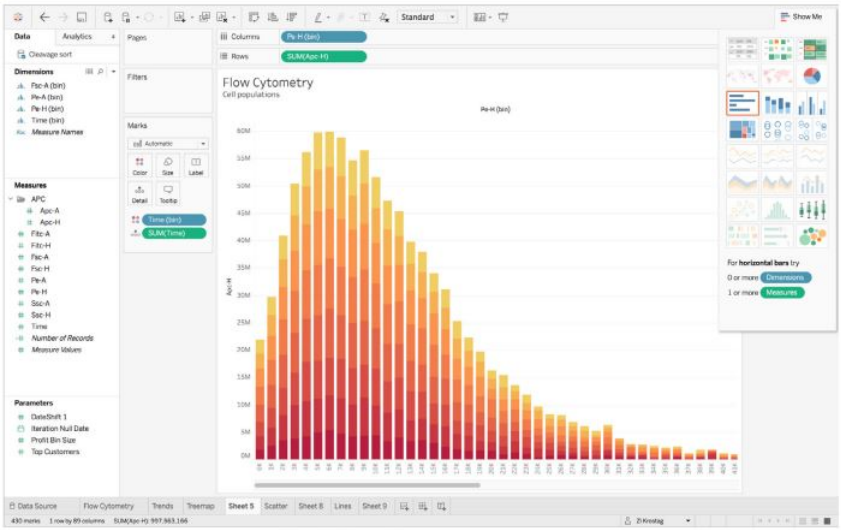


Figura 10 – Exemplo de gráfico: Histograma

Fonte: Tableau (2022).

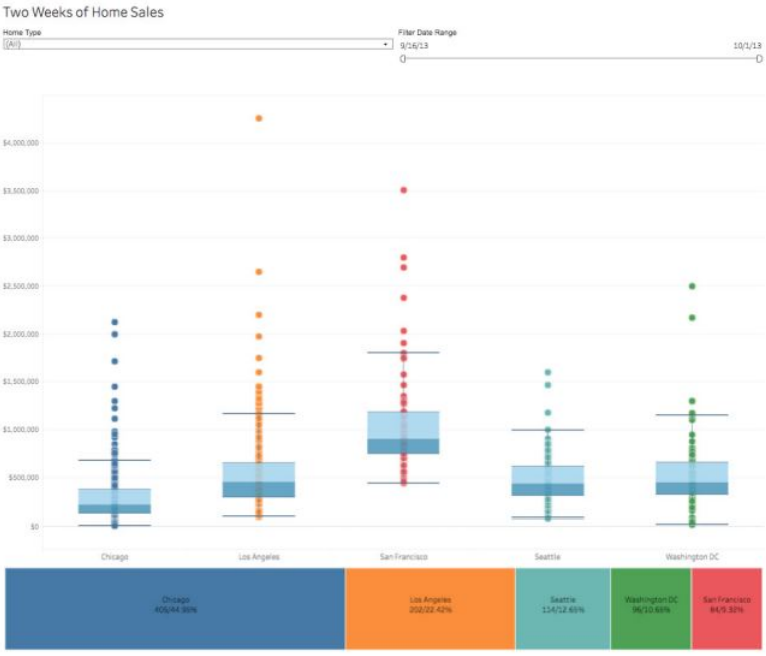


Figura 11 – Exemplo de Gráfico: Whisker-Plot

Fonte: Tableau (2022).

O `Tableau` também permite o uso dessas medidas analíticas, que podem ser: Uma linha Constante, uma linha média, média com quartis, Boxplots, Média e Mediana com intervalo de confiança de 95%, Linha de tendência, *forecasts*, *Clusters*, além de linhas, faixas e distribuição customizadas.

3.6 Conclusões

Visualizações de Dados tem sido usadas como formas de se permitir que conjuntos de dados que podem ser de difícil compreensão – ou seja, que a resposta para a busca do usuário não é facilmente entendida – sejam apresentados de forma adequada (que ele seja capaz de responder às questões de forma clara).

Assim, os objetos complexos deste estudo devem ser disponibilizados de forma a permitir que o usuário tenha a possibilidade de encontrar relações entre diferentes atributos. Além disso, as técnicas de visualização deverão ser exploradas a fim de que se determinem recomendações de quais representações visuais são mais adequadas para a compreensão das correlações de distribuições de espaços de busca, incluindo seus aspectos gráficos.

CAPÍTULO

4

MODELAGEM E VISUALIZAÇÃO DESENVOLVIDAS

Como apresentado na introdução e ao longo das seções desta pesquisa, o objetivo deste trabalho é aumentar a facilidade na identificação de padrões e correlações em estruturas de dados complexos, através da utilização dos conceitos de visualização de dados, similaridade e banco de dados de forma conjunta, oferendo ao especialista do domínio médico a possibilidade de realizar análises em conjunto de dados com correlações de múltiplas variáveis para que, finalmente, ele possa ter *insights* que orientem a tomada de decisão com relação ao uso de fármacos ou procedimentos médicos ou ainda, para que validem alguma hipótese, com o respaldo de avaliações executadas de maneira prática sobre os dados disponíveis.

De fato, a pesquisa desenvolvida percorre e permeia múltiplas questões relacionadas à visualização de dados, arquitetura de banco de dados, descoberta de conhecimento em grandes bases de dados médicos (EHR). Considerando oportuno destacar a participação e importância de cada uma dessas etapas de forma assertiva, compilamos, a seguir, a apresentação macroscópica do conjunto de técnicas apresentadas em um esquema gráfico exibido na figura 12.

Fonte de Dados: Iniciaremos a descrição das etapas pela descrição e caracterização da fonte. Todos os dados utilizados neste trabalho se encontram modelados na `arquitetura EAV (Entidade Atributo Valor)`. Esse tipo de arquitetura preza pela facilidade de inclusão de dados na base, sem a necessidade de estarem padronizados ou normalizados (mas torna a recuperação de conteúdo mais complexa). A justificativa para a adoção desta arquitetura é que a base de dados (1), cedida pelo Instituto do Coração (InCor), segue um `padrão de dados (CDM)` que visa a `interoperabilidade`. Isto é, ela está armazenada em um formato convencionado (OMOP) que permite que outro hospital ou analista, por exemplo, consiga utilizar a base para fazer suas análises combinando esses dados com os de outra instituição. Se não estivessem formatados segundo o OMOP, por exemplo, os valores de colesterol de um hospital não poderiam ser comparados com os de outro, sob pena de se avaliar resultados de calibragens

diferentes ou resultantes de reagentes com sensibilidades variadas. Observe que o OMOP não altera os dados. Ele é um padrão que estabelece a forma como o hospital deve armazenar seus dados e como eles devem se relacionar entre si. Ou seja, a base de dados no formato EAV é o resultado do processamento dos dados originais do hospital InCor, segundo um padrão médico pré-estabelecido, o OMOP-CDM.

Modelagem de arquitetura EAV: Uma vez que esses dados, no padrão OMOP, são gerados (2), um novo desafio, já pré-anunciado surge: Como recuperar informação neste banco de forma intuitiva já que os dados não estão mais em colunas, mas sim, em registros lineares e sem aparente relação. O padrão OMOP, cujo esquema pode ser encontrado nas seções 2 resguardado por suas inegáveis vantagens, é conhecido por sua dificuldade para se conseguir realizar consultas. Os dados deverão ser manipulados em um cliente SQL (neste caso, DBeaver) para que possam ser, em seguida, utilizados em uma ferramenta de visualização ou, até mesmo, em outros tipos de soluções, como fonte de um código na linguagem Python (3).

Visualização de Dados e *Insights*:Ao final dessas etapas, a base volta a se tornar uma base colunar, ou seja, inteligível e pronta para servir de fonte de dados para qualquer outra ferramenta de análise. Contudo, mesmo com a base no formato colunar, os desafios de avaliação de objetos complexos e de correlação de diversas variáveis persiste, o que demanda a aplicação de diversas técnicas (4) para que, finalmente, a descoberta de conhecimento seja de fato, obtida (5).

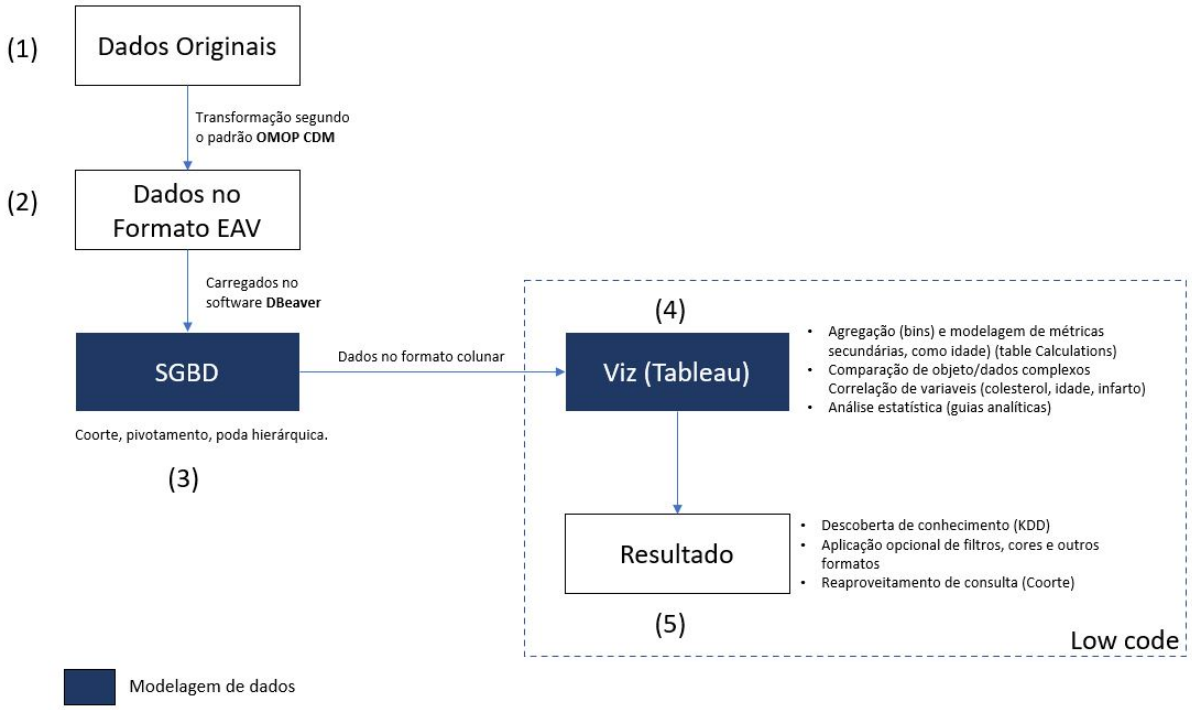


Figura 12 – Organização da pesquisa.

Fonte: Elaborada pelo autor.

Assim, nas seções seguintes, apresentaremos em detalhes as abordagens e métodos utilizados que conduzirão este trabalho ao êxito nos pontos acima destacados.

4.1 A Base de Dados

Utilizamos aqui a base de registros médicos eletrônicos (do inglês, Eletronic Health Records, usualmente chamado EHR (HOERBST; AMMENWERTH, 2010)) do Hospital do Coração (InCor), CDM5, modelada segundo (LIMA *et al.*, 2019).

Nessa base encontram-se 10% dos pacientes do hospital até o ano de 2019. Esses registros foram anonimizados através de um processo de ETL e disponibilizados no formato OMOP, totalizando 94.603 indivíduos. Pela natureza do hospital – referência em atendimentos e em pesquisas para doenças cardíacas, o conjunto de dados concentra grande quantidade de registros de doenças e condições relacionas a essa especialidade médica. Essa informação é relevante para justificar algumas das escolhas das porções de dados para as visualizações e seleções das coortes de dados que serão apresentadas neste trabalho.

Para a exploração desses dados, que serão recuperados da base de dados usando consultas em SQL, foi utilizado um software de desenvolvimento integrado cliente de SQL (*Integrated Development Environment* – IDE) , de código aberto, chamado DBeaver (DBEAVER, 2022) embora ele possa ser substituído por qualquer outro cliente que ofereça os recursos necessários, à escolha do utilizador como o bastante conhecido PGAdmin (PGADMIN, 2022).

Como já destacado, o principal desafio e objetivo deste trabalho consiste em oferecer ao pesquisador ou profissional um caminho onde seja possível visualizar dados que estejam modelados em uma estrutura pouco intuitiva (EAV) e as correlações entre diferentes dimensões de forma prática e eficiente, identificando ainda objetos complexos similares. Para conseguir alcançar esse objetivo, é especialmente importante que o usuário compreenda os principais objetos (tabelas) que compõe esse banco de dados e que seja capaz de manipular os dados para selecionar as coortes mais relevantes para a análise que deseja realizar. As Coortes são seleções bem estabelecidas nos dados e serão amplamente discutidas nas próximas seções deste trabalho.

Como utilizamos a base de dados do Instituto InCor, é evidente que os dados armazenados apresentam um viés de dados relacionado à condições do coração e, portanto, é com relação a esse assunto que está direcionada a análise e as tarefas de visualização de dados apresentadas nos próximos capítulos. O Foco em condições cardíacas apresenta a vantagem de que os resultados empíricos que são obtidos das técnicas de análise estudadas neste trabalho operando sobre dados rotineiros da prática no hospital mostram que o conhecimento considerado bem estabelecido na área podem de fato ser obtidos a partir da análise exploratória dos dados, dando respaldo à validade de conhecimento inédito que pode ser obtido com outras análises equivalentes usando outros dados.

Apesar do foco em condições cardíacas, sendo o termo "condições"referente a qualquer exame, doença ou procedimento envolvendo este órgão ou seus sistemas, como veias, artérias e demais estruturas – a análise de dados é essencialmente exploratória, sem estar baseada em uma hipótese inicial, ou seja, pretende-se analisar os dados sem assumir a existência previa de uma premissa a respeito do resultado que se procura usando premissas de KDD (Knowledge Discovery in Databases) definido como sendo "o processo, não trivial, de extração de informação, implícitas, previamente desconhecidas e úteis, a partir dos dados armazenados em um banco de dados" (FRAWLEY; PIATETSKY-SHAPIRO; MATHEUS, 1991). Contudo, mesmo sendo exploratória, para tratar conjuntos de dados muito grande (*big data*), é de fato, interessante que se aplique algum tipo de restrição ou recorte nos dados para que se tenha otimização de tempo de consulta e de visualização. Assim, nas próximas sessões, serão discutidos alguns dos passos de otimização que foram estudados e que podem ser úteis em diversas situações.

Finalmente, outro aspecto que é explorado nas próximas seções refere-se à apresentação da ferramenta de visualização e da descrição das entidades utilizadas para a análise, além da apresentação de aspectos práticos de uso dos softwares empregados.

4.2 Modelagem de Dados

Cada modelagem de dados deve ser realizada segundo alguns passos e técnicas que consideram também o objetivo de seu uso. Considerando a complexidade e tamanho desta base, tais etapas são fundamentais para assegurar que as análises sejam realizadas através da seleção, caraterização e filtragem dos dados, conforme explicado nas respectivas sessões: `Seleção de Atributo`, `Coortes` e `Poda Hierárquica`. Essas etapas tem por finalidade `restringir` nossa análise a um determinado subconjunto de dados relacionado ao foco que desejamos analisar.

Em seguida, trabalharemos na estrutura dos dados, o que envolve técnicas e operações de manipulação da arquitetura EAV, descritas na sub-sessão `Pivotamento`, que será fundamental para o desenvolvimento das análises e dos resultados apresentada no capitulo 5.

4.2.1 Coorte de Dados

Esta seção apresenta uma definição que será bastante utilizada nas análises e modelagens de dados que se seguirão. Segundo o critério da OHDSI, definimos uma coorte (ou *cohort*, em inglês) como sendo um grupo de indivíduos que atende a determinados critérios estabelecidos. Essa é a forma básica de como os dados são recuperados ou analisados, sempre considerando-se a base de dados no formato OMOP original. Para montar uma coorte, ou seja, obter um recorte considerando certos aspectos nos dados, é necessário que se especifique quais variáveis devem ser analisadas, e isso dependerá unicamente da necessidade do usuário. Observe ainda que a coorte é um recorte virtual e não aloca os dados unicamente em um subconjunto, fazendo

com que um indivíduo possa pertencer a mais de uma coorte, desde que atenda aos critérios determinados pelas mesmas.

As coortes podem ser de dois tipos: A coorte *ruled based* (baseada em regras) e a coorte probabilística, onde a definição dos critérios da seleção será dada por algum tipo de modelo prévio, como um modelo gerado por aprendizado de máquina. Neste trabalho, utilizaremos a *ruled based*, pois ela será suficiente para demonstrar como manipular os dados médicos (EHR) armazenados em uma arquitetura EAV e, em seguida, visualizados por software especializado a fim de induzir descobertas de conhecimento para o usuário.

Vamos deixar pré-estabelecido para as análises e visualizações deste e dos próximos capítulos, que a coorte será escolhida como aquela que recupera dados com as características seguintes, descritas de maneira sintética como:

Incluir todos os indivíduos de qualquer gênero e qualquer idade, que realizaram exames (procedimento de medida) de colesterol e cuja condição inclua infarto. A janela temporal estabelecida é a totalidade dos eventos registrados na base de dados, ou seja, entre os anos de 2003 e 2019.

Essa coorte pode ser obtida recuperando os dados que são delimitados pelas restrições seguintes:

- Evento Inicial da Coorte: Indivíduos que realizaram alguma consulta na modalidade de cardiologia (id = 2000007785).
- Condition Occurrence: Ao menos uma ocorrência.
- Critérios Demográficos: todas as idades, com ambos os gêneros.
- Condition Occurrence Criteria: O tipo de condição é infarto (condition_concept_id = 2000014647) e o tipo de condição é primaria (condition_type_concept_id = ’44786627’) e que tenham realizado exame de colesterol em seu histórico médico.

4.2.2 Realizando poda hierárquica

Nas bases formatadas segundo o padrão OMOP, uma relação hierárquica entre conceitos pode ser estabelecida. O uso de dicionários médicos padronizados com ontologias (nomenclaturas clínicas) que podem ser importadas, como por exemplo o Standard Nomenclature of Medicine (SNOMED), Logical Observation Identifiers Names and Codes (LOINC) são maneiras usuais de se determinar essas relações e classificar doenças e condições segundo uma convenção internacional. Nesta hierarquia, que pode ser compreendida como uma árvore, os conceitos superiores são mais abrangentes, enquanto seus filhos são conceitos mais específicos. Um conceito que está em um nível imediatamente acima ou abaixo na árvore está há um grau de separação. Desta forma, quando nos referirmos a graus de separação, nos referimos à diferença de altura entre os nós da árvore, não discriminado se o nível é descendente ou ascendente.

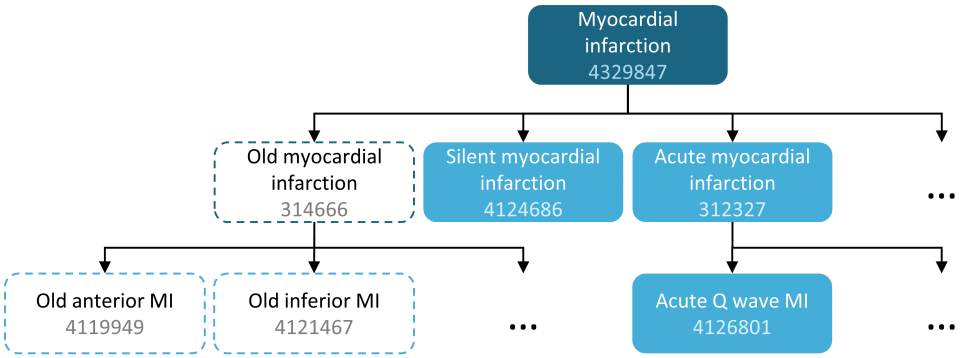


Figura 13 – Hierarquia de Conceitos

Conceito referente a Infarto do Miocárdio e seus descentes.

Fonte: OHDSI (2022).

Esse tipo de distinção pode ser muito valiosa para especialistas do domínio médico, que queiram, por exemplo, examinar apenas os dados referentes a uma parte mais específica para suas análises. Um exemplo ilustrativo pode se referir a um médico que queira avaliar infartos que ocorreram em apenas uma parte especifica do coração. Adicionalmente, se a especificação é feita, do ponto de vista computacional, é ainda bastante útil para aumentar a seletividade da consulta, reduzindo a quantidade de dados retornados, o que pode ser um grande ganho de performance quando se está trabalhando com *big data*. A figura 13 ilustra o exemplo dado.

Apesar de haver a possibilidade de se determinar os descendentes e ascendentes de conceitos no modelo OMOP, é importante lembrar que essas relações precisam ser apontadas pelo desenvolvedor da base. Sem referenciar quais conceitos se relacionam com quais, não é possível utilizar esse tipo de recurso nas consultas e, portanto, nas visualizações e análises de dados. Caso a base que esteja utilizando contenha essa informação, escrevemos as seguintes consultas, que podem ser úteis para indicar algumas formas de se recuperar essas informações:

Exemplo 1: Através dessa consulta, relacionamos, na tabela de conceitos (`Concepts`) os ancestrais dos conceitos (`ancestor_concept_id` da tabela `concept_ancestor`) e, em seguida, os agrupamos e ordenamos, a fim de recuperar os 25 conceitos mais frequentes (linha 7). Aqui, desejamos obter apenas os conceitos relacionados ao domínio dos *Procedures*, que registra todas as manipulações/exames feitas em um paciente (como por exemplo, uma endoscopia) (linha 4). Essa consulta pode pode ser útil para quem deseja compreender e explorar uma base médica, analisando quais os conceitos mais relevantes (com mais ocorrência) em sua base.

Código-fonte 1 – Conceitos mais frequentes na base

```
1: select c2.concept_name from cdm5.concept c2 where c2.concept_id
      IN (
2: select ancestor_concept_id from cdm5.concept c
```

```
3: inner join cdm5.concept_ancestor ca on descendant_concept_id =
     c.concept_id
4: where c.domain_id = 'Procedure' and max_levels_of_separation =
     2
5: group by ancestor_concept_id
6: order by count(ancestor_concept_id) desc
7: limit 25)
```

Exemplo 2: Nesta consulta, utilizamos uma estrutura semelhante ao exemplo anterior. Contudo, aqui, um exemplo mais prático é ilustrado. Suponha que um analista deseja recuperar todos os conceitos que estão relacionados com o ancestral referente à endoscopia (concept_id = 4179713) para sua análise. Ao invés de ser necessário um levantamento nos dicionários médicos e a compreensão uma vasta seleção de conceitos, o mesmo poderia apenas indicar o conceito ancestral e o grau de separação entre os mesmos e obter a listagem dos mais frequentes, conforme o exemplo apresentado na figura 13.

Assim, nesta consulta, recuperamos todos os conceitos cujo ancestral (linha 3), em até 2 níveis de separação, é endoscopia.

Código-fonte 2 – Ancestrais

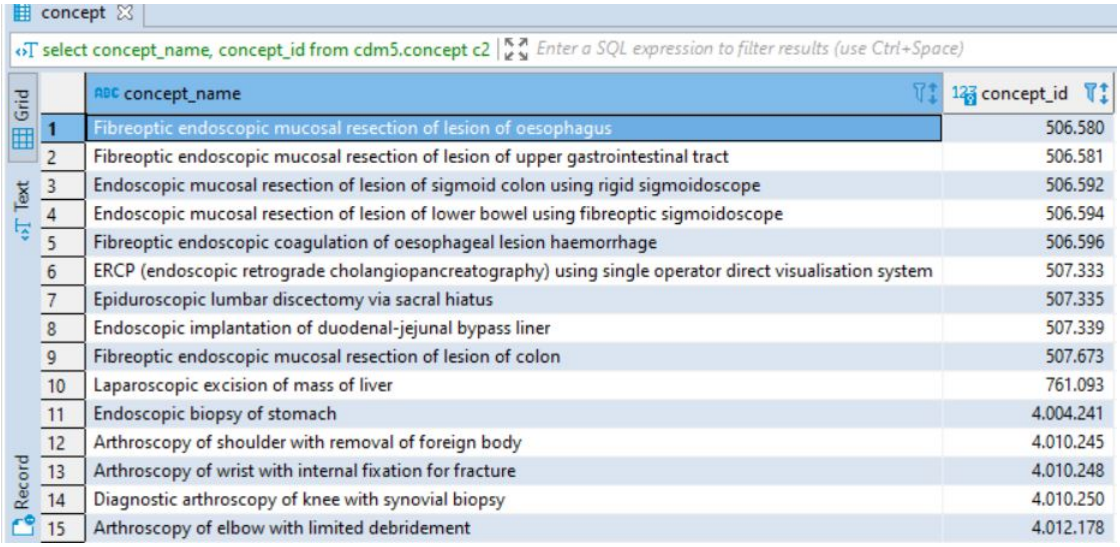
```
1: select concept_name, concept_id from cdm5.concept c2
2: join cdm5.concept_ancestor ca2 on ca2.descendant_concept_id =
     c2.concept_id
3: where ca2.ancestor_concept_id =  '4179713' and
     max_levels_of_separation =2
```

Como já ilustrado na consulta, se o pesquisador desejar limitar o grau de separação entre os conceitos, ele deverá discriminá-lo na consulta indicando a quantidade de graus de forma numérica, assim como na *query* de exemplo acima, que indica o grau 2. O resultado obtido quando executamos a Consulta 2 é o que se mostra na figura 14.

4.2.3 A Operação de Pivotamento

Como colocado nos capítulos iniciais desta monografia, a base de dados que é objeto usado para os exemplos deste mestrado está definida segundo o padrão OMOP-CDM, o qual está formulado segundo uma mescla dos modelos E-R e EAV. Em bases EAV, a descoberta de conhecimento e análise é ineficiente e até mesmo inviável, já que os valores não estão alocados na forma colunar e sim, espalhados em diversas entidades, possuindo redundância de informação e falta de organização. A figura 15 a seguir ilustra e reforça a disposição dos dados nesta arquitetura.

Figura 14 – Resultado da Consulta, Exemplo 2



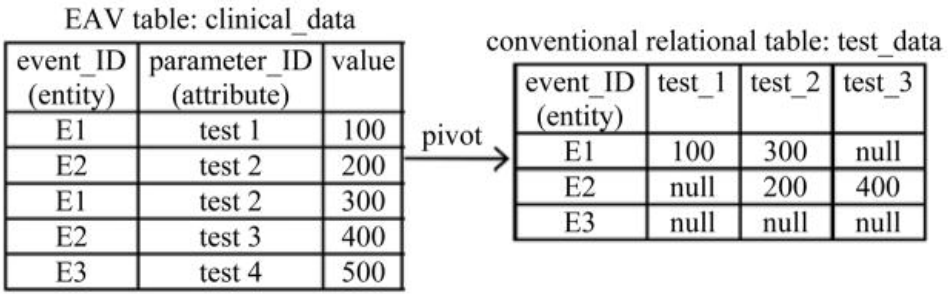
concept

select concept_name, concept_id from cdm5.concept c2 | Enter a SQL expression to filter results (use Ctrl+Space)

Grid | Text | Record

	concept_name	concept_id
1	Fibreoptic endoscopic mucosal resection of lesion of oesophagus	506.580
2	Fibreoptic endoscopic mucosal resection of lesion of upper gastrointestinal tract	506.581
3	Endoscopic mucosal resection of lesion of sigmoid colon using rigid sigmoidoscope	506.592
4	Endoscopic mucosal resection of lesion of lower bowel using fibreoptic sigmoidoscope	506.594
5	Fibreoptic endoscopic coagulation of oesophageal lesion haemorrhage	506.596
6	ERCP (endoscopic retrograde cholangiopancreatography) using single operator direct visualisation system	507.333
7	Epiduroscopic lumbar discectomy via sacral hiatus	507.335
8	Endoscopic implantation of duodenal-jejunal bypass liner	507.339
9	Fibreoptic endoscopic mucosal resection of lesion of colon	507.673
10	Laparoscopic excision of mass of liver	761.093
11	Endoscopic biopsy of stomach	4.004.241
12	Arthroscopy of shoulder with removal of foreign body	4.010.245
13	Arthroscopy of wrist with internal fixation for fracture	4.010.248
14	Diagnostic arthroscopy of knee with synovial biopsy	4.010.250
15	Arthroscopy of elbow with limited debridement	4.012.178

Resultados da consulta número 2 considerando dois graus de separação para o conceito Endoscopia (417913).



EAV table: clinical_data

event_ID (entity)	parameter_ID (attribute)	value
E1	test 1	100
E2	test 2	200
E1	test 2	300
E2	test 3	400
E3	test 4	500

pivot

conventional relational table: test_data

event_ID (entity)	test_1	test_2	test_3
E1	100	300	null
E2	null	200	400
E3	null	null	null

Figura 15 – Operação de Pivotamento.

Fonte: Luo e Frey (2016).

A abordagem mais comum e quase obrigatória para se garantir que a arquitetura EAV possa ser utilizada para análises, não só para que seja inteligível a um humano mas também para que sirva de *input* para outras ferramentas de análise como `python` é utilizar uma operação chamada de `Pivotamento`. O pivotamento consiste em transformar dados que estão alocados em linhas para agrupamentos em colunas. A imagem 15 representa um exemplo conceitual sobre a operação de pivotamento.

Como descrito por (AGRAWAL; SOMANI; XU, 2001) e (LUO; FREY, 2016), a operação de pivotamento pode ser feita de duas formas. A primeira delas, adotada neste trabalho, consiste em se conduzir diversas operações de junção (*joins*) sobre as entidades que se deseja obter informação. A ideia de se realizar *joins* apenas sobre as entidades que se tem interesse no banco aumenta a possibilidade de se obter como resultado uma tabela colunar com baixa esparsidade – a esparsidade deriva do termo relacionado à matrizes esparsas, ou seja, que possuem grande numero de colunas e linhas mas poucas delas com conteúdo diferente de vazio. A baixa esparsidade é desejável, em geral, pois reduz a necessidade de armazenamento e simplifica a estrutura de dados.

Uma escolha atenta das tabelas que armazenam os dados necessários tem papel fundamental na otimização dessa etapa.

Apesar disso, realizar as junções necessárias sobre os dados satisfaz apenas parte do problema uma vez que, apesar dos atributos agora estarem todos distribuídos de forma colunar (ou seja, cada característica da entidade tem seus valores armazenados em uma coluna própria) as linhas ainda mantém uma repetição de elementos que não oferecem ao usuário uma forma direta de avaliar o resultado encontrado apos a operação de *outer pivot*.

Assim, o passo seguinte para a modelagem de uma arquitetura EAV é a `agregação dos atributos` para que haja apenas um registro para cada objeto e se possa criar estruturas como chaves primárias, fundamentais para análises de dados. Aqui, surge mais um trunfo da análise de dados baseada em softwares de visualização - O usuário pode, neste momento, utilizar o resultado desse pivotamento diretamente no software, por exemplo no `Tableau` e as agregações e modelagens que serão necessárias para se fazer a descoberta de conhecimento das mesmas sem precisar despender ainda mais esforço de modelagem de banco de dados. Ou seja, parte do processo de pivotamento de bases EAV que tradicionalmente precisam ser feitos nos SGBDs podem ser feitos de maneira visual no `Tableau`, conforme será explorado no próximo capítulo

4.3 Conclusões

Bases de dados que armazenam informações de EHR em geral são construídas considerando-se uma arquitetura contra-intuitiva e de difícil recuperação de conteúdo, chamada EAV. Essa escolha se dá principalmente pela necessidade de interoperabilidade entre diferentes instituições e para que se consiga integrar os dados provenientes de diversas bases de dados que seguem estruturas e semânticas distintas, para viabilizar as análises sobre as mesmas, diminuindo a interferência externa sobre as modelagens de dados que caso contrário seriam necessárias. Apresentamos algumas abordagens bastante úteis como a definição de coortes, a poda hierárquica e o pivotamento de dados. Para essa última etapa, apresentamos uma nova abordagem, utilizando uma solução híbrida entre SGBDs e ferramentas de visualização a fim de se obter a agregação dos registros que viabiliza aproveitar o interesse do usuário em cada consulta e a subsequente análise de dados e geração de *insights.*

O próximo capítulo será dedicado à explanação sobre última etapa de tratamento de dados, bem como o resultado final de visualizações realizadas sobre essa base.

CAPÍTULO

5

RESULTADOS

O que se busca em um trabalho de visualização? Ora, que se possa avaliar informações de um conjunto de dados que não são, a princípio, interpretáveis de maneira imediata e simples a um cérebro humano. Trata-se, portanto, na boa visualização, da capacidade acurada para se escolher, dentro de uma ampla gama de opções de regras de sumarizações e agrupamentos, cores e relações, aquelas que têm o maior potencial de enfatizar a resposta para uma pergunta ou de exibir relações não antes notadas em uma coleção de dados.

Desta forma, encontrar uma forma de visualização de dados organizados em um modelo EAV pode ser ainda mais desafiante, já que os atributos de um objeto fogem de um modelo normalizado que siga uma modelagem relacional mais bem estruturada, que não estão diretamente e claramente relacionados ou mesmo na mesma entidade, fazendo com que a identificação de padrões e correlações entre dimensões e atributos necessite de ainda mais análises para ser identificada. O que iremos apresentar, a seguir, consiste no resultado das abordagens e técnica apresentadas no capítulo anterior para que, com elas, possamos conseguir atingir o objetivo deste trabalho: `Visualizar correlações e identificar` ***`insights`*** `sobre EHR a partir de um banco que registra dados sob o modelo EAV`.

Dentro dessas premissas, faz-se importante destacar neste trabalho que nosso objetivo é fornecer uma forma de se visualizar tais relacionamentos, limitando-se a oferecer soluções tecnológicas mas não tendo a intenção de prescrever nenhum tratamento ou avaliação médica. Esse passo caberá aos profissionais de saúde no domínio desta ciência, visando facilitar suas tarefas com o uso dos resultados obtidos pelas técnicas aqui descritas.

A seguir serão exemplificadas e explicadas as técnicas de modelagem de dados apresentadas no capitulo anterior, a fim de se obter os dados especificados em uma coorte definida pelos profissionais de saúde, representadas em estruturas de dados adequadas à visualização, utilizando diferentes procedimentos de consulta sobre um banco de dados modelado segundo o modelo EAV. Em seguida, será apresentada a segunda parte deste tópico, relacionada à modela-

gem desses dados dentro do software Tableau, incluindo a manipulação de dados e a criação de métricas segundo as técnicas e conceitos apresentados no capítulo 3 e conceitos como a similaridade, discutida no capítulo 2. Finalmente, como resultado, apresentaremos os resultados e visualizações que podem ser obtidas.

5.1 Exploração da Base

Iniciaremos a análise demostrando o processo de exploração da base, utilizando os conceitos apresentados no capítulo anterior aplicados na base descrita anteriormente (base InCor). Neste trabalho, realizamos as consultas diretamente em um cliente SQL, utilizando o software DBeaver.

Relembrando a metodologia utilizada, descrita com mais detalhes no capítulo 4, onde a coorte utilizada foi definida, a modelagem inicial envolve a realização de diversos *joins* (pivotamento) utilizando todas as tabelas necessárias para compor o conjunto de dados (filtragem) que será consumido na análise.

5.1.1 Entidades Principais

O início da modelagem se dá pela determinação de quais deverão ser os objetos presentes na base de trabalho a ser visualizada. Considerando a coorte estabelecida, é necessário que se compreenda o conteúdo das entidades do padrão OMOP descritas a seguir.

- Condition Occurrence:

  A tabela Condition Occurence (em português, Condições das Ocorrências) armazena todos os registros de um indivíduo, extraídos de uma fonte de dados primária, como é o caso por exemplo, de um diagnóstico armazenado no Sistema de Informações Hospitalares primário da instituição que fornece o dado, e convertido para a representação OMOP por um processo automático controlado pela mesma instituição. Note-se que uma ocorrência é registrada mesmo que uma condição semelhante já tenha sido registrada anteriormente. As condições são extraídas de diversos tipos de fontes de dados dos sistemas de informação hospitalares, possivelmente de diversos subsistemas e que empregam tipos de formatação diferentes em diferentes instituições – por exemplo, contemplando processos laboratoriais e equipamentos de coleta de fabricantes diversos. Portanto, é necessário atenção quando se deseja avaliar diferentes condições em um mesmo momento ou análise.

  Os principais atributos desse tipo de entidade são o identificador unívoco daquela ocorrência, o identificador (Person_Id), que referencia qual indivíduo teve essa condição registrada, os atributos referentes ao registro de data, atributos que classificam a condição em tipos (seguindo a normalização indicada em Condition_Type_Concept_Id), e aquele que referencia em qual visita de um profissional de saúde essa ocorrência foi registrada.

Para a maioria das análises deste e de outros trabalhos, esses atributos são os mais utilizados, pois eles permitem responder a uma ampla variedade de perguntas, como por exemplo, *Por quanto tempo perdurou uma condição?*, *Qual é a distribuição das condições entre os pacientes da coorte?* ou ainda, *Será que essa condição é sazonal?*. A lista completa dos atributos desta tabela e o significado dos atributos categóricos pode ser encontrada na documentação do OMOP-CDM (CDM5-4doc, 2022).

- `Procedure Occurrence`:

  A tabela `Procedure_Occurence` (em português, Ocorrências de Procedimentos) contém os registros das atividades e processos realizados em/por um paciente a fim de se obter um diagnostico ou alguma evidencia médica. Um procedimento pode ser entendido como a execução de um exame (mas não seu resultado), uma checagem de sinais vitais como por exemplo uma medida de temperatura ou qualquer outra coleta de dados em um exame, um procedimento médico ou de enfermaria como por exemplo uma endoscopia ou uma desobstrução arterial, etc. Os dados que usualmente são conhecidos como Prontuário Eletrônico do Paciente – PEP (em inglês *Electronic Health Record* – EHR) estão majoritariamente armazenados em registros desta entidade.

  Os principais atributos desta entidade são o identificador unívoco de identificação do precedimento, o paciente que se submeteu a esse procedimento, o identificador do procedimento, datas e o provedor (a referência à instituição fonte) destes dados.

- `Person`:

  Essa entidade armazena as informações demográficas, que identificam cada paciente. As informações mais pertinentes contidas nesta entidade é o identificador unívoco do paciente, gênero, data de nascimento, etnia e localização. Vale ressaltar que nenhum dado que permita a identificação do paciente, como nome ou qualquer documento de identidade deve ser registrado nesta entidade.

- `Measurement`:

  A entidade `Measurement` (ou medidas, em português). Esta tabela armazena dados numéricos e categóricos que foram gerados através de exames ou testagem de um paciente, como os resultados de exames laboratoriais, achados patológicos, sinais vitais etc. Os principais atributos dessa entidade são o identificador unívoco da medida, o identificador do paciente, atributos relacionados a datas, valores absolutos e a faixa de variação máxima e mínima de valores admitidos como normais para o exame específico a que essa medida se refere (range_low e range_high) além de um campo que informa qual a unidade de medida foi utilizada. Veja que uma mesma medida, executada segundo dois procedimentos diferentes pode ter valores e e faixa de valores normais diferentes.

5.1.2 Pivotamento

Uma vez que temos a coorte estabelecida no capítulo anterior, definida como *Todos os indivíduos de qualquer gênero e qualquer idade, que realizaram exames (procedimento) de colesterol e cuja condição é infarto sobre toda a base de dados sem restrição de período* e, portanto, definidas quais devem ser as entidades utilizadas baseados no conteúdo que cada entidade armazena, devemos seguir com o pivotamento dos dados. Ou seja, as informações que estão armazenadas em linhas devem ser transladadas para que ocupem a posição correspondente de uma coluna na modelagem final com os respectivos fatores de conversão aplicados para que os valores possam ser comparados.

Como descrito no capítulo 4, a seguinte consulta foi escrita para realizar o *outer pivot*. Nela, as entidades listadas acima foram unidas pelas chaves determinadas pelo padrão OMOP-CDM V5.4 – que é atualmente, a versão totalmente suportada mais recente. Assim, as junções combinam, através das chaves determinadas no modelo para as entidades acima listadas, segundo o modelo de relacionamento ilustrado no capítulo 2. O resultado da consulta está apresentado na figura 16. Observe-se que as linhas de 1 a 6 referem-se a um mesmo indivíduo, o que é esperado visto que iremos realizar a etapa de agregação na visualização, facilitando ainda mais a operação para um analista da área de saúde, que possivelmente tenha pouca familiaridade com a linguagem SQL.

```
1: select p.person_id, p.year_of_birth, p.gender_source_value,
2: m.measurement_date, m.measurement_source_value, m.
     value_as_number, m.range_low,
3: m.range_high,  co.visit_occurrence_id as condition_visit_id,
4: co.condition_concept_id, condition_start_date,
5: condition_end_date, condition_type_concept_id,
6: (select c2.concept_name from cdm5.concept c2
7: where co.condition_concept_id = c2.concept_id ) as
     condition_concept_name,
8: po.procedure_concept_id, po.procedure_date, c.concept_name as
     procedure_name
9: from cdm5.condition_occurrence co
10: join cdm5.procedure_occurrence po on po.person_id = co.
     person_id
11: join cdm5.concept c on c.concept_id = po.procedure_concept_id
12: join cdm5.person p on p.person_id = po.person_id
13: join cdm5.measurement m on m.person_id = po.person_id
14: --where po.procedure_concept_id = '2000004800' -- id de
     eletrocardiograma como procedimento (exemplo extra)
15: where po.procedure_concept_id = '2000007785' --id de consulta
     de cardiologia como procedimento (definição de coorte)
```

```
16: and co.condition_type_concept_id = '44786627' and
17: co.condition_concept_id = '2000014647' -- id referente a condiç
      ão de infarto (definição de coorte)
18: and measurement_source_value = 'VAPA/COLESTEROL' (definição de
      coorte)
19: --and m.value_as_number > 250 --valor considerado alto (exemplo
       extra: especificação de faixa de valores na definição da
      coorte)
20:
21: --44786627  Primary Condition
```

Essa consulta inclui as condições de filtragem da coorte definida, que foi modelada na consulta através dos seguintes comandos:

1. Atendendo ao requisito da coorte - Infarto (condição satisfeita na linha 17);
2. Atendendo ao requisito da coorte - Exame de colesterol (satisfeita em linha 18);
3. Atendendo ao requisito do paciente ter sido atendido em uma consulta de cardiologia (satisfeita em linha 15).

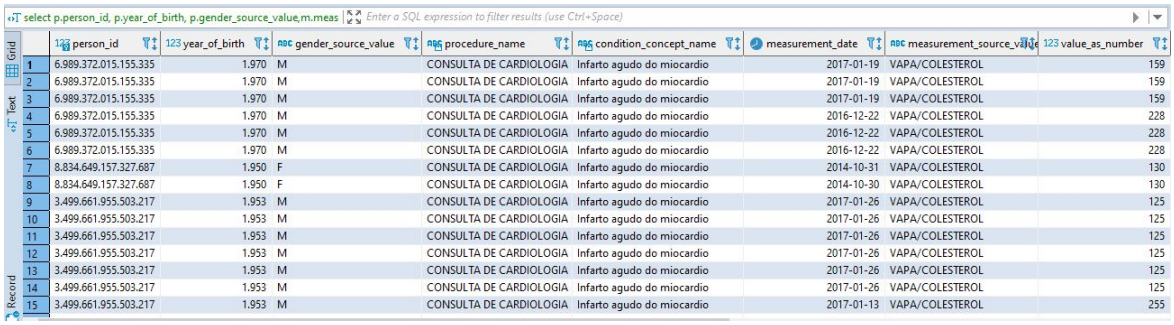
Recuperando dados demográficos para análise:

1. **Idade e data de nascimento e gênero do paciente:** Recuperados na definição do Select, linha 1;
2. **janela temporal:** Sem restrição de data aplicada.

Aqui colocamos alguns exemplos adicionais que podem ser úteis para o analista de dados recuperar informações/modelar características de sua coorte: Considerar apenas quem teve colesterol acima de uma faixa determinada (linha 19); Considerar quem fez algum tipo de procedimento (exemplo eletrocardiograma) (linha 14).

5.2 Visualização de correlações multidimensionais e Objetos Complexos

Para analisar como as variáveis de medidas escalares como colesterol, idade e infarto se correlacionam (variáveis que estão presentes na coorte), é necessária uma análise multifatorial e tridimensional. Além disso, compreender o formato da distribuição da ocorrência de infartos considerando essas variáveis torna ainda mais complexa a análise de dados e a compreensão dos resultados de forma clara pelo analista de dados.



select p.person_id, p.year_of_birth, p.gender_source_value,m.meas | Enter a SQL expression to filter results (use Ctrl+Space)

	person_id	year_of_birth	gender_source_value	procedure_name	condition_concept_name	measurement_date	measurement_source_value	value_as_number
1	6.989.372.015.155.335	1.970	M	CONSULTA DE CARDIOLOGIA	Infarto agudo do miocardio	2017-01-19	VAPA/COLESTEROL	159
2	6.989.372.015.155.335	1.970	M	CONSULTA DE CARDIOLOGIA	Infarto agudo do miocardio	2017-01-19	VAPA/COLESTEROL	159
3	6.989.372.015.155.335	1.970	M	CONSULTA DE CARDIOLOGIA	Infarto agudo do miocardio	2017-01-19	VAPA/COLESTEROL	159
4	6.989.372.015.155.335	1.970	M	CONSULTA DE CARDIOLOGIA	Infarto agudo do miocardio	2016-12-22	VAPA/COLESTEROL	228
5	6.989.372.015.155.335	1.970	M	CONSULTA DE CARDIOLOGIA	Infarto agudo do miocardio	2016-12-22	VAPA/COLESTEROL	228
6	6.989.372.015.155.335	1.970	M	CONSULTA DE CARDIOLOGIA	Infarto agudo do miocardio	2016-12-22	VAPA/COLESTEROL	228
7	8.834.649.157.327.687	1.950	F	CONSULTA DE CARDIOLOGIA	Infarto agudo do miocardio	2014-10-31	VAPA/COLESTEROL	130
8	8.834.649.157.327.687	1.950	F	CONSULTA DE CARDIOLOGIA	Infarto agudo do miocardio	2014-10-30	VAPA/COLESTEROL	130
9	3.499.661.955.503.217	1.953	M	CONSULTA DE CARDIOLOGIA	Infarto agudo do miocardio	2017-01-26	VAPA/COLESTEROL	125
10	3.499.661.955.503.217	1.953	M	CONSULTA DE CARDIOLOGIA	Infarto agudo do miocardio	2017-01-26	VAPA/COLESTEROL	125
11	3.499.661.955.503.217	1.953	M	CONSULTA DE CARDIOLOGIA	Infarto agudo do miocardio	2017-01-26	VAPA/COLESTEROL	125
12	3.499.661.955.503.217	1.953	M	CONSULTA DE CARDIOLOGIA	Infarto agudo do miocardio	2017-01-26	VAPA/COLESTEROL	125
13	3.499.661.955.503.217	1.953	M	CONSULTA DE CARDIOLOGIA	Infarto agudo do miocardio	2017-01-26	VAPA/COLESTEROL	125
14	3.499.661.955.503.217	1.953	M	CONSULTA DE CARDIOLOGIA	Infarto agudo do miocardio	2017-01-26	VAPA/COLESTEROL	125
15	3.499.661.955.503.217	1.953	M	CONSULTA DE CARDIOLOGIA	Infarto agudo do miocardio	2017-01-13	VAPA/COLESTEROL	255

Figura 16 – As 15 primeiras linhas resultantes da consulta principal.

Observe que parte dos atributos não é exibida como forma de preservar a leitura da imagem em uma escada adequada.

Fonte: Elaborada pelo autor.

Tradicionalmente, análises envolvendo mais de um vínculo de correlação não são tão simples de se produzir, principalmente quando se deseja também encontrar a curva de frequência de ocorrência de eventos (neste caso, infartos) dentro de uma população. Já dentro do campo da visualização, uma abordagem direta poderia considerar um elemento tridimensional. Contudo, esse tipo de recurso é reconhecidamente pouco eficaz no campo de Data Viz, já que frequentemente causa distorção nas dimensões.

Desta forma, deve-se responder à tarefa de como se pode produzir um visual de compreensão simples e que traga consigo ainda referencias estatísticas, tais como a distribuição de valores e medidas de tendência central para o evento de interesse (infarto). Apresentaremos, a seguir, uma abordagem de como é possível atingir esse objetivo de forma simples mas efetiva.

5.2.1 Relacionando Colesterol e Idade de Condições de Infarto

Considerando a natureza dos dados da base utilizada (Hospital InCor), para este trabalho foram escolhidas as análises referentes a doenças e condições do coração. A aplicação dos conceitos e técnicas descritos nos capítulos anteriores buscam satisfazer a três objetivos principais:

- Capacidade de identificar relações valiosas para as análises entre os dados;
- Capacidade de relacionar dados em uma visualização utilizando uma base EAV;
- Facilidade de uso por pessoas que não dominam técnicas de mineração de Dados e KDD.

Para se relacionar os dados de infarto, idade e colesterol, uma série de junções devem ser feitas na base de dados. Essas operações podem ser feitas tanto em um cliente SQL quanto diretamente dentro do software de visualização (tal como o `Tableau`, utilizado nesta dissertação). A escolha entre esses dois caminhos deve levar em consideração principalmente a quantidade de

dados a ser relacionada, já que o software de visualização tem, tipicamente, uma capacidade de processar dados em volume menor do que um sistema projetado para para lidar com consultas SQL e deve considerar a sofisticação da escrita das consultas para atender as necessidades especificas do projeto, como por exemplo, a necessidade de utilizar, por exemplo, funções amostrais como SAMPLE e necessidade de uso de funções definidas pelo usuário. A necessidade de ter dados transacionais, que podem estar sendo coletados em tempo real (ou quase tempo real) pode ser uma das motivações para que o usuário prefira fazer as junções e transformações diretamente no `Tableau`, usando a interface e a documentação descritas na seção 4.4.1.

5.2.2 Operando o `Tableau`

5.2.2.1 Inserindo os dados

Uma vez que os dados tenham sido extraídos e modelados, conforme explicado nas seções anteriores, pode-se iniciar a etapa de tratamento e visualização dentro do `Tableau`. Para isso, a base de dados foi importada diretamente dentro da plataforma – como explicado anteriormente, a conexão direta com a base é também possível.

Iniciaremos essa modelagem apresentando a interface inicial do software `Tableau`. A primeira tela mostra os dados que serão carregados e dispõe de ferramentas para o *parsing* (divisão sintática dos dados). O *parsing* pode ser feito tanto de forma automática quanto através da especificação do usuário a respeito da divisão dos dados.

Aqui, também é possível determinar qual o tipo do dado – ou seja, determinar se o tipo de determinada coluna é de tipo `int`, `texto (string)`, `booleano`, `data/tempo` etc. Essa determinação é importante porque a modelagem que poderá ser aplicada sobre determinada dimensão ou métrica depende do seu tipo. Por exemplo, somas e outras operações matemáticas só podem ser realizadas sobre atributos numéricos e alguns tipos de divisões de tempo só podem ser aplicadas sobre campos do tipo data. Na figura 17 está representada a tela inicial com os dados da coorte gerada, bem como a indicação e explicação das seções que compõe essa tela.

5.2.2.2 Modelando os dados

A etapa seguinte consiste em se utilizar os visuais pré-determinados do `Tableau` para a análise de dados. Para o dataset gerado pela coorte, portanto, temos como objetivo relacionar 3 atributos: Valor do exame de colesterol, idade e ocorrência de infarto. Para identificar relações entre as três variáveis consideradas, serão necessárias as etapas:

- Cálculo da idade do paciente na ocasião do infarto;
- Criação de faixas de similaridade de condições de infarto;
- Inserção de guias analíticas.

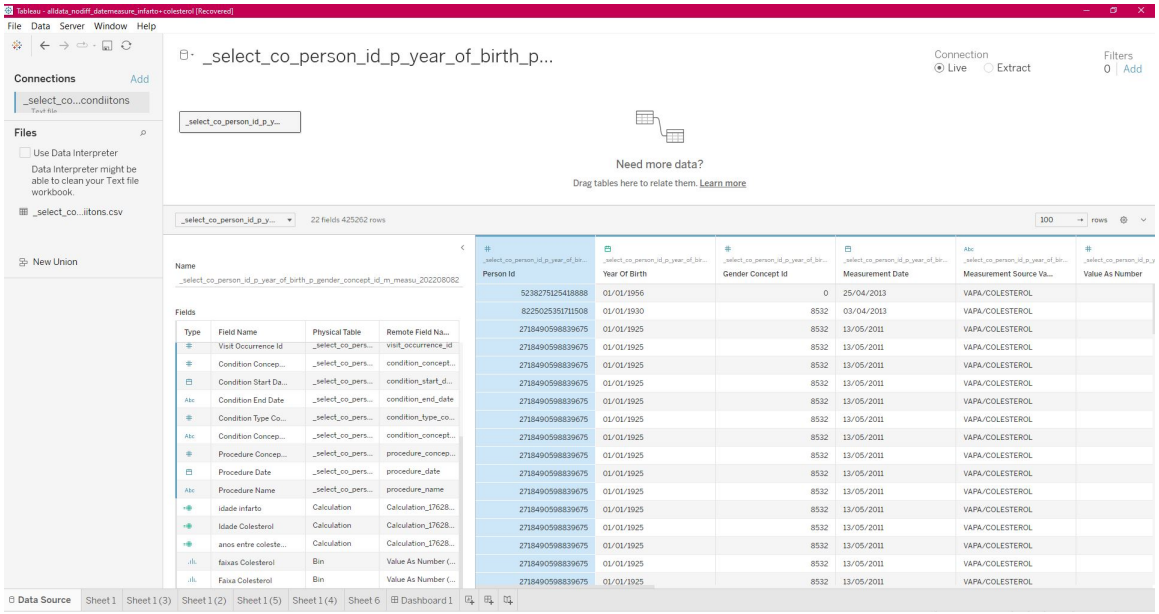


Figura 17 – Tableau - Tela de carregamento dos Dados gerados pela coorte.

Fonte: Elaborada pelo autor.

Precisamos, portanto, calcular a idade do paciente no momento do infarto. Note que essa não é uma das informações que o banco de dados possui, é necessário calcular essa medida. Note ainda que `Tableau` está preparado para esse tipo de modelagem, conforme descrito no capitulo 3. Assim, uma *Table Calculation* é necessária. Para isso, selecionamos a Guia *Analysis* e em seguida, a opção *Create Calculated Field*. Na tela que surge, escrevemos seguindo a sintaxe proprietária do `Tableau` a especificação do cálculo que é necessário, conforme mostrado na figura 18:

Figura 18 – Tableau - *Calculated Field* (idade).

Fonte: Elaborada pelo autor.

Agora, é possível criar um visual contendo a distribuição das idades dos infarto, conforme apresentado na figura 28. Um primeiro resultado de visualização surge agora: `É possível verificar que a idade onde mais ocorrem infartos é 64 anos.`

Na etapa seguinte, desejamos adicionar a `terceira dimensão` para avaliar a relação entre a idade, a frequência de infartos e o nível de colesterol no sangue que o indivíduo apresentou durante seu histórico de saúde. Contudo, os valores de colesterol apresentam uma variação muito grande. Os valores assumidos para o colesterol vão de uma faixa mínima e não possuem limites,

sendo ainda graduados em unidades de décimos de valor, o que faria com que um gráfico ficasse muito poluído e praticamente impossível de ser lido ou avaliado. Para confirmarmos essa afirmação, o `Tableau` oferece a possibilidade de verificar a descrição de determinado atributo e, para os valores de colesterol, temos 209 valores diferentes na coorte que adicionamos como fonte de dados. Assim, a figura 19 mostra como o usuário pode obter a descrição do atributo, o que pode ser bastante útil para a escolha de como ou qual visual utilizar e a figura 20 mostra como ficaria a visualização destes dados, seguindo a distribuição de infartos por idade e por faixa de colesterol se utilizássemos esse atributo diretamente, sem tratamento. Observe que obter algum *insight* dessa visualização é difícil e não intuitivo.

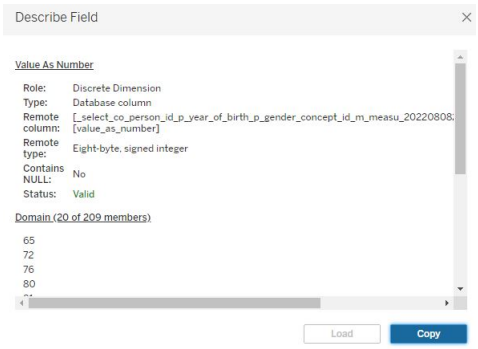


Figura 19 – Tableau - Descrição do Atributo Colesterol.

Fonte: Elaborada pelo autor.

Uma abordagem adequada é considerar que há `similaridade` entre os pacientes que estão dentro de uma faixa de colesterol. Conforme apresentado como uma das bases deste trabalho, a utilização de similaridade para a análise de EHR é muito valiosa já que uma condição (como infarto) ou valores de determinados exames, como por exemplo, colesterol, varia dentro de uma grande faixa de valores (*range*). Apesar disso, a diferença de poucas unidades entre os valores pode representar um mesmo quadro ou condição, sendo, portanto, condições ou `pacientes similares`. Como apresentado no capítulo 2, a similaridade pode ser calculada por funções de distância ou coeficientes de similaridade. Utilizaremos neste trabalho a função de distância euclidiana e determinaremos um raio para uma consulta por abrangência (*range query*) para agrupar os valores dentro de faixas.

Como a determinação das faixas deve levar em consideração um conhecimento do domínio de estudo – neste caso a medicina, estabelecemos valores que podem ser revistos por médicos ou outros especialistas. Assim, determinam-se as faixas de valores e a classificação dos pacientes similares para faixas de 50mg/dl de colesterol. Para isso, iremos utilizar outra função do `Tableau`, conforme as imagens 21 e 22.

Com os *bins* criados segundo a abrangência determinada, temos o visual apresentado na figura 23. Agora, a visualização dos dados fica muito mais fácil e a avaliação das variáveis escolhidas pode ser feita de maneira mais intuitiva. Podemos ainda melhorar a capacidade de análise desses dados, inserindo guias analíticas à visualização. Para isso, utilizaremos a

Figura 20 – Tableau - Distribuição da ocorrência de infarto por idade e por valor de colesterol

Observe-se que a visualização sem tratamento da dimensão colesterol não favorece *insights* relevantes.

Fonte: Elaborada pelo autor.

guia *Analytics*, no painel localizado na lateral esquerda do `Tableau` 24. Utilizamos como configuração de nível de análise os painéis (se escolhêssemos o valor de tabela, o `Tableau` iria realizar os cálculos considerando todos os valores e não cada faixa de forma independente). Assim, adicionando os rótulos aos dados, chegamos à forma final do visual desejado, conforme mostrado na figura 25.

5.2.3 Outros Exemplos

A fim de ilustrar outras possibilidades de uso da ferramenta e da exploração de dados de objetos complexos, foram criadas outras visualizações. Os dados utilizados pertencem à mesma coorte e foram modelados de forma semelhante às etapas descritas no exemplo completo anterior.

Na análise realizada, poderíamos querer compreender qual a incidência de infartos relacionando a idade, o nível de colesterol `e o gênero do paciente`. A figura 26 mostra este resultado.

A figura 27 mostra a relação entre infarto e idade utilizando outro tipo de visualização. Note que este tipo de visualização é capaz de destacar rapidamente os pontos de atenção e relevância dos dados, diferentemente do que aconteceria se os mesmos estivessem dispostos em uma tabela.

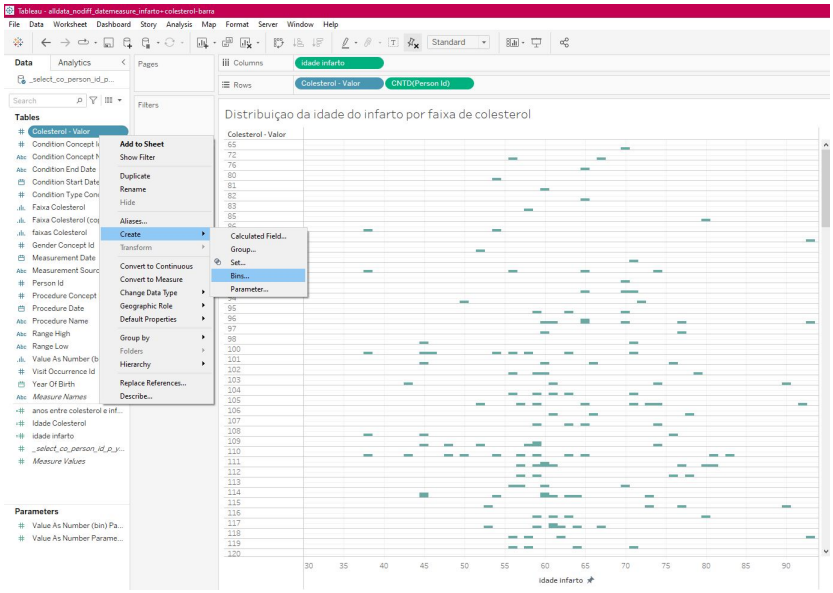


Figura 21 – `Tableau` - A ferramenta permite a criação de *bins* para o agrupamento dos dados.

Fonte: Elaborada pelo autor.

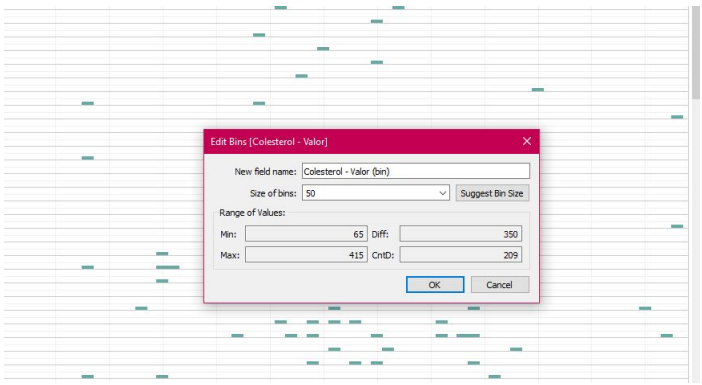


Figura 22 – `Tableau` - Janela para configuração das características dos *bin*.

Aqui, selecionamos o valor 50 como a abrangência (*range*) da função de similaridade.

Fonte: Elaborada pelo autor.

5.3 Resultados

5.3.1 Análise das visualizações

Antes de iniciar a análise deste resultado, faz-se necessário relembrar que o objetivo deste trabalho é oferecer um caminho viável para viabilizar a análise de dados médicos armazenados em prontuários eletrônicos (EHR) armazenados sob a arquitetura do modelo EAV, sem a necessidade de se utilizar técnicas avançadas de mineração, as quais muitas vezes estão distante do dia a dia médico em atividade num hospital. Este trabalho, contudo, não tem por objetivo fazer qualquer avaliação sobre a utilização de medicamentos, procedimentos ou condições médicas, deixando esta avaliação para os especialistas deste domínio. Uma vez feitas as ressalvas, podemos, portanto, considerar o resultado dessa visualização.

Figura 23 – `Tableau` - Distribuição dos dados de infarto por idade e faixa de colesterol.

Fonte: Elaborada pelo autor.

Na figura 23, cada uma das faixas horizontais representa uma faixa de colesterol e as barras verticais representam um histograma com a distribuição das idades dos indivíduos que sofreram infarto. Cada uma das faixas possui a indicação do valor médio para a idade onde o infarto aconteceu, bem como a quantidade média de indivíduos com esse valor. Já a região preenchida com a cor cinza mostra o intervalo de 95% de precisão dos dados. Com isso, é fácil de se observar que há uma correlação entre a idade do paciente e o nível de colesterol no seu sangue. Ou seja, é possível identificar que quanto maior a faixa de colesterol, mais cedo esse paciente provavelmente sofrerá um infarto.

As faixas com valor de colesterol acima de 250 mg/dl mostram uma idade superior mas a confiabilidade para essa amostra se torna mais reduzida, dada a menor quantidade de indivíduos nesta faixa de valores.

Se for utilizada apenas uma das variáveis, como por exemplo, a idade, para se compreender quando é mais provável que um paciente tenha um infarto (ou seja, a definição de um grupo de risco), sem se considerar seu nível de colesterol, poderiam ser classificados como baixo risco pacientes com alta chance de infartar. A figura 25 mostra que a idade média de infarto para pacientes com nível de colesterol acima de 200 mg/dl é de 59.4 anos mas, se considerarmos apenas a dimensão de idade, a média de idade para uma pessoa ter um infarto é 5 anos maior (64 anos), conforme apresentado na figura 28.

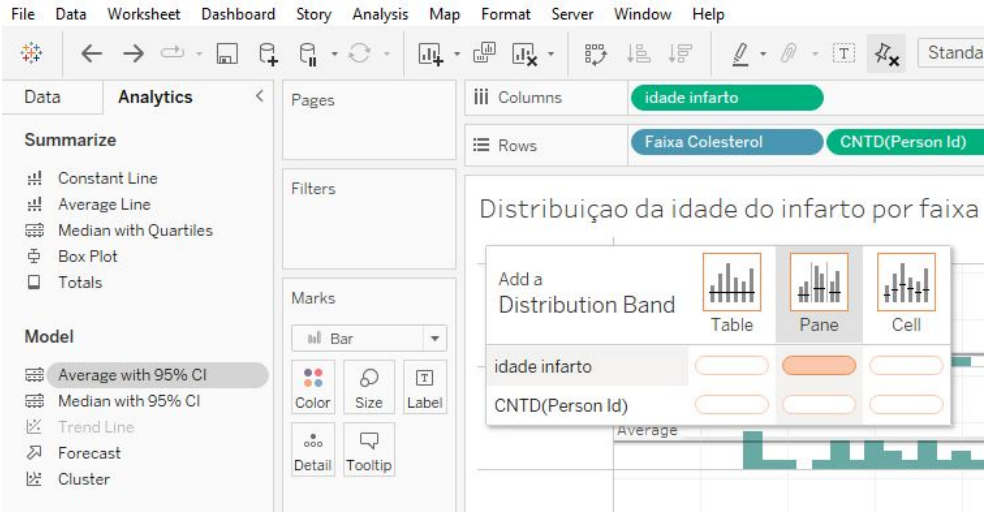


Figura 24 – `Tableau` - Inserção de Guias Analíticas.

É possível escolher o tipo de análise desejada.

Fonte: Elaborada pelo autor.

5.3.2 Conclusões

O resultado das visualizações comprovam, de maneira empírica e prática, avaliando sobre os dados reais das atividades diárias do hospital, alguns dos conhecimentos que são correntes entre os profissionais da área. Isso tem especial significado por se tratar de análises executadas sobre dados colhidos ao longo de um amplo período de tempo e sobre muitos casos (pacientes).

Também se destaca neste capítulo, a demonstração da capacidade de se realizar comparações e agrupamentos entre objetos complexos, como pacientes, descritos por diversas características de forma intuitiva, mesmo quando suas características estão armazenadas em um modelo de dados não normalizado (EAV), gerando *insights* acionáveis e inteligíveis por profissionais de outras áreas, o que dificilmente seria obtido por técnicas específicas e tradicionais para consultas de similaridade.

Finalmente, apontamos como resultados adequados apresentados nesta seção, a capacidade das visualizações serem ajustadas e re-modeladas através de filtros, novas disposições visuais e cálculos de métricas derivadas, sem a necessidade de se operar novamente sobre os SGBDs, permitindo o reaproveitamento de código e de processos de análise bem-sucedidas, atendendo de forma responsiva à eventuais novas necessidades ou para responder à diferentes perguntas que evoluem ao longo do tempo.

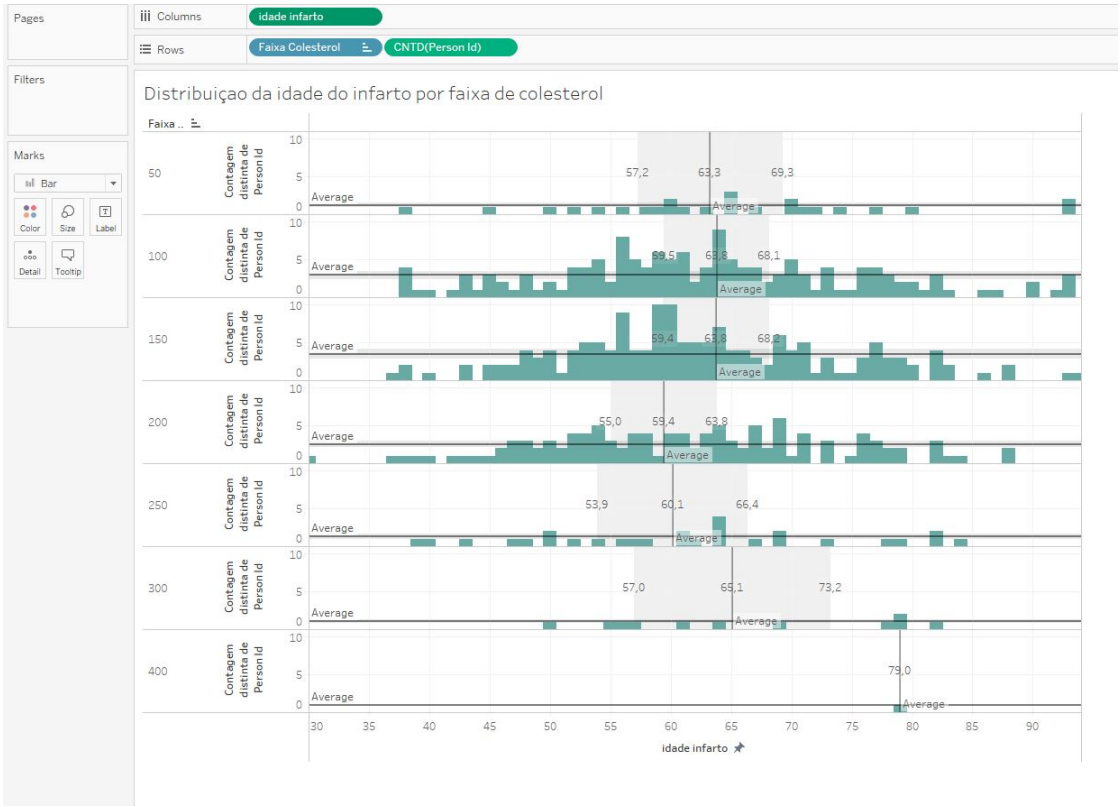


Figura 25 – Tableau - Visual final.

Visual com guias analíticas, cálculos de métricas e agrupamento de valores.

Fonte: Elaborada pelo autor.

Figura 26 – `Tableau` - Ocorrências de Infarto por faixa de colesterol e idade e gênero.

A cor vermelha representa pacientes do sexo feminino quanto a cor laranja representa o masculino. Os dados em azul representam indivíduos sem informação sobre gênero na base..

Fonte: Elaborada pelo autor.

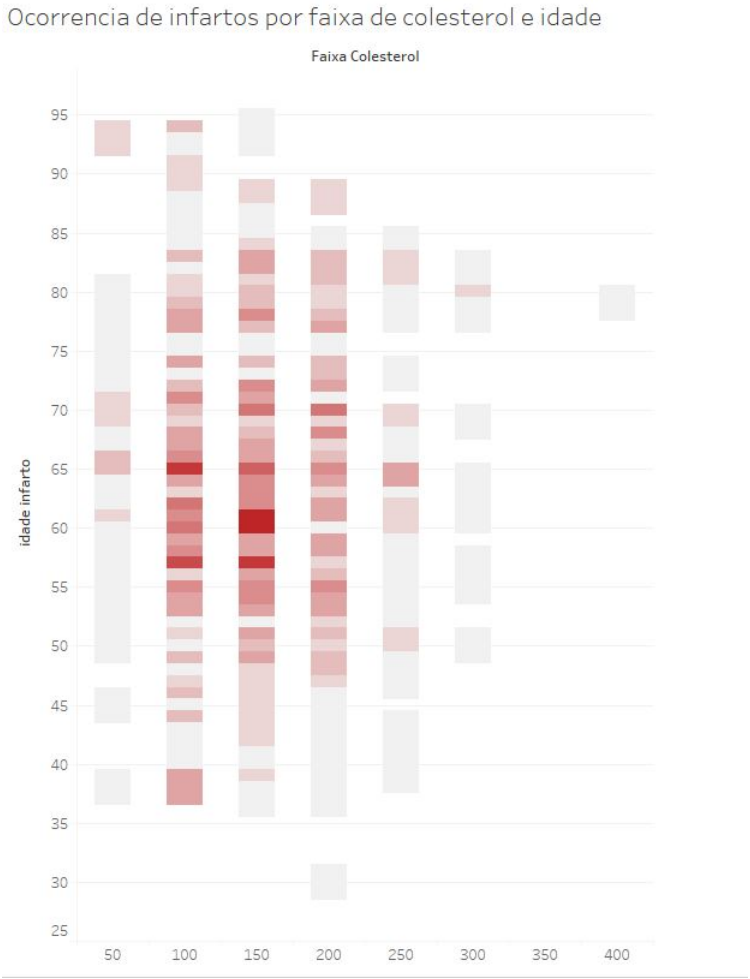


Figura 27 – Tableau - Ocorrências de Infarto por faixa de colesterol e idade.

Observe que quanto mais forte a cor, maior a quantidade de pessoas infartadas.

Fonte: Elaborada pelo autor.

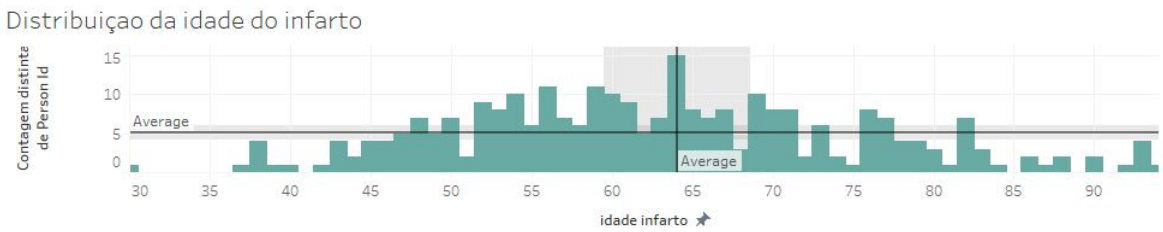


Figura 28 – Média Geral - Idade Infarto

CAPÍTULO

6

CONCLUSÕES E TRABALHOS FUTUROS

6.1 Conclusões

Utilizando como convenção a mesma estrutura de pesquisa que foi apresentada no capítulo 4, destacamos aqui, como executados, os desafios elencados na metodologia ilustrada em 12.

Modelagem de arquitetura EAV (1, 2 e 3): Analisando a extensa documentação do Modelo de dados padrão OMOP-CDM com seus domínios, relações, dicionários e formas de recuperação de dados específicas, processos de refinamento como a poda hierárquica, em conjunto com a análise e apoiando o estudo da base do Instituto do Coração (InCor) e dos demais dados armazenados, demostramos como escrever consultas que recuperam os dados referentes à coorte escolhida e determinada, levantando os registros médicos (EHR) de pacientes e de suas ocorrências de infarto, consultas médicas de cardiologia e exames de colesterol e formatando os dados em estruturas adequadas aos processos de análise visual.

Visualização de Dados (4 e 5): Em seguida, com os dados já representados em formato colunar, ilustramos a importação da base gerada no software `Tableau`, onde foram realizadas modelagens sobre os dados, a fim de se obter métricas derivadas, agregações e filtros que são necessários para a compreensão da coorte. Além disso, uma vez que refinamentos e manipulações sobre a `base` tenham sido feitas, podem ser aplicados ajustes e outras técnicas de visualização aos resultados gerados, garantindo um resultado final (25)que possa viabilizar ao analista uma melhor avaliação do paciente, considerando os diferentes tipos de dados que existem sobre ele de maneira integrada, que passa então a poder ser tratado como um objeto complexo, descrito por diversas variáveis, cada uma de diferentes maneiras, de forma simples e prática e ainda estabelecer faixas de similaridade entre eles, que podem ser facilmente ajustadas, de acordo com a necessidade ou critérios médicos.

6.1.1 Vantagens e Conquistas deste trabalho

Como vantagens de se utilizar a metodologia proposta para a descoberta de conhecimento em bases heterogêneas, destacamos:

- Modelagem de dados (etapa de pivotamento e agrupamento), que foi feita diretamente dentro do software `Tableau`, não necessitando de modelagens em SGBDs, o que pode facilitar a democratização desse tipo de análise por parte de pessoas que não dominam a analise de dados ou a gestão do sistemas de gerenciamento de banco de dados e das bases de dados armazenadas.

- Apesar de haver a necessidade do uso de SGBDS para a formatação armazenagem das tabelas-base, todas as etapas de descoberta de conhecimento podem ser realizadas em ambiente *no/low code*, o que pode ser benéfico para equipes ou estruturas com divisões entre a área de modelagem e a área de inteligencia da instituição de saúde.

- A abordagem para a descoberta de conhecimento em *bigdata* de conjuntos de dados provenientes dos Registros Eletrônicos dos Pacientes em bases mantidas em modelagens EAV, que são reconhecidamente difíceis de navegar por conta de sua estrutura pouco intuitiva. Mostramos como a análise de correlação entre três variáveis pode ser feita, utilizando uma interface amigável e gerando resultados inteligíveis mesmo para pessoal não especializado em Tecnologia da Informação.

- Verificamos que a capacidade de obter painéis dentro de um sistema otimizado e dedicado para geração, armazenamento, manutenção e apresentação de *insights* sobre dados pode se mostrar uma grande vantagem frente a geração de imagens e gráficos feitos em sistemas de domínio público, não comerciais, como aqueles utilizando bibliotecas como `seaborn` em `python` ou `ggplot` em R, linguagens tradicionais para análises de dados. O uso de software de visualização, em especial, o `Tableau`, oferece a possibilidade do analista não precisar despender tempo organizando e atualizando de forma mais manual e direta todas as análises que em muitas vezes, são recorrentes, além de não depender de um local ou sistema para disponibilizá-lo para os tomadores de decisão que, na maioria das situações não dominam as ferramentas e linguagens usadas.

6.2 Trabalhos Futuros

Este trabalho de mestrado apresentou o resultado prático da aplicação de técnicas de visualização de dados utilizando os padrões de visualização pré-estabelecidos dentro da ferramenta `Tableau`, considerando as limitações de tempo deste mestrado. Apesar disso, considerando as técnicas e conceitos apresentados, o estudo sobre a criação de um tipo específico de visual, que permitisse ao usuário aplicar diversos tipos de funções de similaridade sobre a base carregada, oferecendo uma forma espontânea de encontrar e relacionar pacientes ou condições similares dentro de uma coorte estabelecida se apresentaria como abordagem valiosa, oferecendo uma

análise de dados ainda mais poderosa de forma intuitiva e agnóstica ao domínio do conhecimento, que pode ser aplicada a outras áreas de atividade.

O desenvolvimento deste tipo de ferramenta, se acompanhada e guiada por um especialista do domínio médico, pode ser feita considerando os aspectos mais relevantes e desafiadores enfrentados por esses profissionais no dia a dia e, portanto, a sugerimos que a interface inter-especialidades (computação e medicina) seja aplicada. Assim, afigura-se que diversos trabalhos podem dar continuidade a este trabalho, dentre os quais destacamos aqueles descritos a seguir, os quais deverão ser executados em cooperação com especialistas da área da saúde.

Construção de uma Biblioteca de Padrões de Visualização – Estabelecer protocolos padrões de projeto para a geração de *templates* para visualização de dados, que podem ser classificados pelo objetivo das informações obtidas, segundo o interesse do usuário.

Construção padronizada de Coortes – Definir coleções de estruturas de dados para coortes que se destinem a diferentes padrões de visualização. Esta atividade é especialmente interessante devido à possibilidade que dados proveniente de diversas estruturas básicas dos dados podem levar a dados que contemplem a obtenção de dados úteis para analisar situações correlatas.

Definição de modelos de interoperabilidade entre instituições parceiras – Apesar da modelagem EAV permitir a representação de dados provenientes de diversas fontes, a criação de processos de integração de dados provenientes de pares de instituições pode agilizar a geração de coortes específicas para processos de visualização pré-definidos.

Definição de outras operações de comparação e busca por similaridade – Processos de integração baseados em quantidade de objetos similares e não apenas em faixas de variação pode ser interessantes, especialmente em situações em que existem muitos dados muito similares.

Processos de agrupamento e classificação baseados em similaridade – Desenvolver processos de visualização de dados que tenham por objetivo identificar diferentes classes de padrões em coortes ou executar levantamentos baseados em estatísticas de agrupamento identificados em coortes.

6.3 Publicações

Utilizando os conceitos deste trabalho, tal como a análise de bases heterogêneas, a descoberta de conhecimento explorando a correlação de variáveis e o agrupamento similar de dados e visualização, foi desenvolvida, em conjunto com demais colegas do grupo de pesquisa um artigo aceito e publicado na revista Journal of Data, Information and Management (JDIM, Springer).

O artigo, intitulado *"Analysis of Enem's attendants between 2012 and 2017 using a clustering approach"* (em português Análise dos estudantes do ENEM usando uma abordagem

de agrupamento) (LIMA *et al.*, 2020) apresentou formas de tratamento de bases heterogêneas – já que os dados analisados eram de diferentes anos, com coleta de dados e objetivos de procedimentos de análise díspares, e possuíam diferentes formatos e valores diferentes para modelos de pontuação diferentes. Em seguida, foi desenvolvida uma técnica de análise de similaridade utilizando o método de `k-Means` para dividir os dados em grupos, para a partir dos grupos criar visualizações que destacassem as semelhanças e diferenças entre eles.

A análise mostrou discrepâncias de performance no exame entre diferentes áreas do Brasil, bem como situações entre os alunos relacionadas ao tipo de ensino (público ou privado) e outras variáveis demográficas, como a participação de estudantes com necessidades especiais. O resultado é bastante relevante para se avaliar as políticas públicas de educação e, eventualmente, oferecer um respaldo baseado em dados para o desenvolvimento de políticas públicas para a melhoria ou ajustes nas próximas edições do exame, tendo potencial concreto de aplicação imediata na sociedade brasileira.

REFERÊNCIAS

AGGARWAL, A.; SHARMA, S.; SINGH, K.; SINGH, H.; KUMAR, S. A new approach for effective retrieval and indexing of medical images. **Biomedical Signal Processing and Control**, v. 50, p. 10–34, 2019. Citado na página 26.

AGRAWAL, R.; SOMANI, A.; XU, Y. Storage and querying of e-commerce data. In: **VLDB**. [S.l.: s.n.], 2001. v. 1, p. 149–158. Citado na página 56.

ALTSCHUL, S. F.; BOGUSKI, M. S.; GISH, W.; WOOTTON, J. C. Issues in searching molecular sequence databases. **Nature Genetics**, v. 6, p. 119–129, 1994. Citado na página 28.

ANDERSON, T. W. **An Introduction to Multivariate Statistical Analysis**. New York: [s.n.], 1984. Citado na página 28.

AZEVEDO-MARQUES, P. M. d.; MENCATTINI, A.; SALMERI, M.; RANGAYYAN, R. M. **Medical Image Analysis and Informatics: Computer-Aided Diagnosis and Therapy**. 1. ed. London: CRC Press, Taylor & Francis Group, 2017. v. 1. Citado na página 41.

BARIONI, M. C. N.; KASTER, D. d. S.; RAZENTE, H. L.; TRAINA, A. J. M.; JR, C. T. Querying multimedia data by similarity in relational DBMS. In: YAN, L.; MA, Z. (Ed.). **Advanced Database Query Systems: Techniques, Applications and Technologies**. Hershey, NY, USA: IGI Global, 2010. p. 323–359. Citado na página 30.

BRANCATI, N.; CAMASTRA, F. Analysis of similarity measurements in cbir using clustered tamura features for biomedical images. In: PIETRO, G. d.; GALLO, L.; HOWLETT, R. J.; JAIN, L. C. (Ed.). **Intelligent Interactive Multimedia Systems and Services 2016**. Puerto de la Cruz, Tenerife, Spain: Springer International Publishing, 2016. p. 1–10. Citado na página 26.

BRUNO, N.; CHAUDHURI, S. Interactive physical design tuning. In: **ICDE, International Conference on Data Engineering**. [S.l.]: IEEE Computer Society, 2010. p. 1161–1164. Citado na página 21.

BUDIKOVA, P.; BATKO, M.; ZEZULA, P. **Query Language for Complex Similarity Queries**. 2012. 23, CoRR abs/1204.1185 p. Citado na página 30.

CDM5-4doc. 2022. https://github.com/OHDSI/CommonDataModel/releases/tag/v5.4.0. Accessed: 2022-11-21. Citado nas páginas 35 e 61.

CDM54. 2022. <https://ohdsi.github.io/CommonDataModel/>. Accessed: 2022-26-11. Citado na página 35.

CHEN, L.; GAO, Y.; LI, X.; JENSEN, C. S.; CHEN, G. Efficient metric indexing for similarity search. In: **2015 IEEE 31st International Conference on Data Engineering**. [S.l.: s.n.], 2015. p. 591–602. Citado na página 27.

CHEN, R. S.; NADKARNI, P.; MARENCO, L.; LEVIN, F.; ERDOS, J.; MILLER, P. L. Exploring performance issues for a clinical database organized using an entity-attribute-value representation. **Journal of the American Medical Informatics Association**, v. 7, n. 5, p. 475–487, 2000. Citado na página 34.

CODD, E. F. A relational model of data for large shared data banks. **Communications of the ACM (CACM)**, v. 13, n. 6, p. 377–387, 1970. Citado na página 21.

DATE, C. J. **SQL and Relational Theory - How to Write Accurate SQL Code**. [S.l.]: O'Reilly Media, 2009. Citado nas páginas 21 e 28.

DBEAVER. 2022. <https://dbeaver.io/>. Accessed: 2022-27-11. Citado na página 51.

DEZA, M. M.; DEZA, E. **Encyclopedia of Distances**. 4th. ed. Heidelberg: Springer, 2016. Citado nas páginas 27 e 28.

DÖLLER, M.; TOUS, R.; TEMMERMANS, F.; YOON, K.; PARK, J.-H.; KIM, Y.; STEGMAIER, F.; DELGADO, J. Jpeg's jpsearch standard: Harmonizing image management and search. **IEEE Multimedia**, v. 20, n. 4, p. 38–48, 2013. Citado na página 26.

ELMASRI, R.; NAVATHE, S. **Fundamentals of Database Systems**. [S.l.]: Pearson, 2010. Citado na página 32.

ELMASRI, R.; NAVATHE, S. B. **Sistemas de Banco de Dados - Tradução da** 6^a **Edição Americana**. 6. ed. [S.l.]: Pearson Education - Br, 2011. Citado na página 28.

FRAWLEY, W. J.; PIATETSKY-SHAPIRO, G.; MATHEUS, C. J. Knowledge discovery in databases: An overview. In: PIATETSKY-SHAPIRO, G.; FRAWLEY, W. J. (Ed.). [S.l.]: AAAI/MIT Press, 1991. p. 1–27. Citado na página 52.

GARCÍA-MOLINA, H.; ULLMAN, J. D.; WIDOW, J. **Database Systems: The Complete Book**. 2. ed. [S.l.]: Pearson Education, 2011. Citado na página 28.

GARTNER. 2022. <https://www.gartner.com/doc/reprints?id=1-29HD7D53&ct=220323&st=sb>. Accessed: 2022-27-11. Citado na página 42.

GÜLD, M. O.; THIES, C.; FISCHER, B.; LEHMANN, T. M. A generic concept for the implementation of medical image retrieval systems. **International Journal of Medical Informatics (IJMI)**, v. 76, p. 252, 2007. Citado na página 27.

GUSFIELD, D. Core string edits, alignments, and dynamic programming. In: **Algorithms on Strings, Trees, and Sequences - Computer Science and Computational Biology**. Cambridge, MA: Cambridge University Press, 1997. v. 1, p. 215–253. Citado na página 28.

HAAK, D.; PAGE, C.-E.; DESERNO, T. M. A survey of DICOM viewer software to integrate clinical research and medical imaging. **J. Digital Imaging**, v. 29, n. 2, p. 206–215, 2016. Citado na página 41.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining - Concepts and Techniques, 3rd Edition**. 3st edition. ed. New York: Morgan Kaufmann Publishers, 2012. Citado na página 26.

HANRAHAN, P. Vizql: a language for query, analysis and visualization. In: **Proceedings of the 2006 ACM SIGMOD international conference on Management of data**. [S.l.: s.n.], 2006. p. 721–721. Citado na página 42.

HERZOG, T. N.; SCHEUREN, F. J.; WINKLER, W. E. **Data Quality and Record Linkage Techniques**. [S.l.]: Springer, 2007. Citado na página 28.

HOERBST, A.; AMMENWERTH, E. Electronic health records. **Methods of information in medicine**, Schattauer GmbH, v. 49, n. 04, p. 320–336, 2010. Citado na página 51.

JAGADISH, H. V.; MENDELZON, A. O.; MILO, T. Similarity-based queries. In: **ACM Symp. on Principles of Database Systems (PODS)**. San Jose, CA: ACM Press, 1995. p. 36–45. Citado na página 28.

JR., C. T.; MORIYAMA, A.; ROCHA, G. M. d.; CORDEIRO, R. L. F.; CIFERRI, C. D. d. A.; TRAINA, A. J. M. The SimilarQL framework: similarity queries in plain SQL. In: **The 34th ACM/SIGAPP Symposium on Applied Computing, SAC 2019**. Limassol, Cyprus: ACM, 2019. p. 468–471. Citado na página 30.

KEOGH, E. J. Exact indexing of dynamic time warping. In: BRESSAN, S.; CHAUDHRI, A. B.; LEE, M.-L.; YU, J. X.; LACROIX, Z. (Ed.). **International Conference on Very Large Databases (VLDB)**. Hong Kong, China: Springer Verlag, 2002. (Lecture Notes in Computer Science 2590), p. 406–417. Citado na página 28.

KHALID, N. E. A.; YUSOFF, M.; KAMARU-ZAMAN, E. A.; KAMSANI, I. I. Multidimensional data medical dataset using interactive visualization star coordinate technique. **Procedia Computer Science**, v. 42, p. 247 – 254, 2014. ISSN 1877-0509. Medical and Rehabilitation Robotics and Instrumentation (MRRI2013). Disponível em: <http://www.sciencedirect.com/science/article/pii/S1877050914014975>. Citado na página 41.

LIMA, A.; FLOREZ, A.; LESCANO, A.; NOVAES, J.; MARTINS, N.; JUNIOR, C. T.; SOUSA, E.; JÚNIOR, J. R.; CORDEIRO, R. Analysis of enem's attendants between 2012 and 2017 using a clustering approach. **Journal of Information and Data Management**, v. 11, n. 2, 2020. Citado na página 78.

LIMA, D. M.; RODRIGUES-JR, J. F.; TRAINA, A. J. M.; PIRES, F. A.; GUTIERREZ, M. A. Transforming two decades of ePR data to OMOP CDM for clinical research. **Stud. Health Technol. Inform.**, v. 264, p. 233–237, ago. 2019. Citado nas páginas 36 e 51.

LIMA, E. L. **Espaços Métricos**. [S.l.]: Instituto de Matemática Pura e Aplicada, 1993. Citado na página 27.

LIU, Y.; ZHANG, D.; LU, G.; MA, W.-Y. A survey of content-based image retrieval with high-level semantics. **Pattern Recognition Letters**, v. 40, p. 262 – 282, 2007. Citado na página 26.

LÖPER, D.; KLETTKE, M.; BRUDER, I.; HEUER, A. Enabling flexible integration of healthcare information using the entity-attribute-value storage model. **Health Information Science and Systems**, v. 1, n. 1, p. 9, 2013. Citado na página 34.

LU, W.; HOU, J.; YAN, Y.; ZHANG, M.; DU, X.; MOSCIBRODA, T. MSQL: efficient similarity search in metric spaces using SQL. **VLDB J.**, v. 26, n. 6, p. 829–854, 2017. Citado na página 30.

LUO, G.; FREY, L. J. Efficient execution methods of pivoting for bulk extraction of entity-attribute-value-modeled data. **IEEE J. Biomed. Health Inform.**, Institute of Electrical and Electronics Engineers (IEEE), v. 20, n. 2, p. 644–654, mar. 2016. Citado na página 56.

MALIK, R.; KIM, S.; JIN, X.; RAMACHANDRAN, C.; HAN, J.; GUPTA, I.; NAHRSTEDT, K. Mlr-index: An index structure for fast and scalable similarity search in high dimensions. In: WINSLETT, M. (Ed.). **21st International Conference on Scientific and Statistical Database Management, SSDBM 2009**. New Orleans, LA: Springer, 2009. (Lecture Notes in Computer Science, v. 5566), p. 167–184. Citado na página 29.

MELTON, J.; SIMON, A. R. **SQL:1999 Understanding Relational Language Components**. 1. ed. [S.l.]: Morgan Kaufmann, 2002. (The Morgan Kaufmann series in Data Management Systems). Citado na página 21.

NADKARNI, P. M.; MARENCO, L.; CHEN, R.; SKOUFOS, E.; SHEPHERD, G.; MILLER, P. Organization of heterogeneous scientific data using the eav/cr representation. **Journal of the American Medical Informatics Association**, v. 6, n. 6, p. 478–493, 1999. Citado na página 33.

NAVATHE, S. B. R. E. **Fundamentals of database systems**. Hoboken, New Jersey: Pearson, 2017. Citado na página 33.

NIEDERMAYER, J. **Complex queries and complex data: challenges in similarity search**. Tese (Doutorado) — Fakultat fur Mathematik, Informatik und Statistik der Ludwig-Maximilians-Universitat, 2015. Citado na página 27.

OHDSI. 2022. https://ohdsi.github.io/TheBookOfOhdsi/Cohorts.html. Accessed: 2022-11-21. Citado na página 54.

PAPAPETROU, P.; ATHITSOS, V.; POTAMIAS, M.; KOLLIOS, G.; GUNOPULOS, D. Embedding-based subsequence matching in time-series databases. **ACM Transactions on Database Systems (TODS)**, v. 36, n. 3, p. 1–39, 2011. Citado na página 28.

PGADMIN. 2022. <https://www.pgadmin.org/>. Accessed: 2022-27-11. Citado na página 51.

POLA, I. R. V.; TRAINA, A. J. M.; JR, C. T. Easing the dimensionality curse by stretching metric spaces. In: WINSLETT, M. (Ed.). **21st International Conference on Scientific and Statistical Database Management, SSDBM 2009**. New Orleans, LA: Springer, 2009. (Lecture Notes in Computer Science, v. 5566), p. 417–434. Citado na página 29.

PUTTMANN, D.; KEIZER, N. D.; CORNET, R.; ZWAN, E. V. D.; BAKHSHI-RAIEZ, F. FAIRifying a quality registry using OMOP CDM: Challenges and solutions. **Stud. Health Technol. Inform.**, v. 294, p. 367–371, maio 2022. Citado na página 36.

SAJJAD, M.; ULLAH, A.; AHMAD, J.; ABBAS, N.; RHO, S.; BAIK, S. W. Integrating salient colors with rotational invariant texture features for image representation in retrieval systems. **Multimedia Tools Appl.**, v. 77, n. 4, p. 4769–4789, 2018. Citado na página 26.

SAMET, H. **Foundations of Multidimensional and metric Data Structures**. San Francisco, CA: Morgan Kaufmann Publishers, 2006. Citado na página 29.

SILVA, Y. N.; ALY, A. M.; AREF, W. G.; LARSON, P.-A. Simdb: a similarity-aware database system. In: **Proceedings of the 2010 international conference on Management of data**. Indianapolis, Indiana, USA: ACM, 2010. p. 1243–1246. Citado na página 30.

SILVA, Y. N.; AREF, W. G.; ALI, M. H. The similarity join database operator. In: LI, F.; MORO, M. M.; GHANDEHARIZADEH, S.; HARITSA, J. R.; WEIKUM, G.; CAREY, M. J.; CASATI, F.; CHANG, E. Y.; MANOLESCU, I.; MEHROTRA, S.; DAYAL, U.; TSOTRAS, V. J. (Ed.).

International Conference on Data Engineering, ICDE. Long Beach, California, USA: IEEE Computer Society, 2010. p. 892–903. Citado na página 30.

SIMPSON, M. S.; RAHMAN, M.; SINGHAL, S.; DEMNER-FUSHMAN, D.; ANTANI, S. K.; THOMA, G. R. Text- and content-based approaches to image modality detection and retrieval for the imageclef 2010 medical retrieval track. In: BRASCHLER, M.; HARMAN, D.; PIANTA, E. (Ed.). [S.l.]: CEUR-WS.org, 2010. p. 1–10. Citado na página 25.

STOLTE, C.; TANG, D.; HANRAHAN, P. Polaris: A system for query, analysis, and visualization of multidimensional relational databases. **IEEE Transactions on Visualization and Computer Graphics**, IEEE, v. 8, n. 1, p. 52–65, 2002. Citado nas páginas 42 e 43.

TABLEAU. 2022. <https://www.tableau.com/pt-br>. Accessed: 2022-27-11. Citado nas páginas 42 e 46.

TEKLI, J.; CHBEIR, R.; TRAINA, A. J. M.; JR, C. T. XML document-grammar comparison: Related problems and applications. **Central European Journal of Computer Science**, v. 1, n. 1, p. 117–136, 2011. Citado na página 28.

TELEA, A. C. **Data Vizualization Principles and Practice**. [S.l.]: Taylor & Francis Group, 2015. Citado nas páginas 38, 39 e 40.

THOMAS, P. C. W. J. Visual analytics. **IEEE Computer Graphics and Applications**, 2004. Citado na página 41.

TRAINA, A. J. M.; JR, C. T.; BALAN, A. G. R.; RIBEIRO, M. X.; BUGATTI, P. H.; WATANABE, C. Y. V.; MARQUES, P. M. d. A. Feature extraction and selection for decision making over medical images. In: DESERNO, T. M. (Ed.). **Biomedical Image Processing - Methods and Applications**. [S.l.]: Springer-Verlag, 2010. p. 197–223. Citado nas páginas 26, 29 e 30.

TRAINA, A. J. M.; JR, C. T.; CIFERRI, C. D. d. A.; RIBEIRO, M. X.; MARQUES, P. M. d. A. How to cope with the performance gap in content-based image retrieval systems. **International Journal of Healthcare Information Systems and Informatics (IJHISI)**, v. 4, n. 1, p. 47–67, 2009. Citado na página 27.

WELTER, P.; FISCHER, B.; GÜNTHER, R. W.; LEHMANN), T. M. Deserno (né. Generic integration of content-based image retrieval in computer-aided diagnosis. **Computer Methods and Programs in Biomedicine**, 2012. Citado na página 26.

WILLIAMS, J. G.; SOCHATS, K. M.; MORSE, E. Visualization. **Annual Review of Information Science and Technology (ARIST) 30**, p. 161– 207., 1995. Citado na página 37.

WU, L.; JIN, R.; JAIN, A. K. Tag completion for image retrieval. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 35, n. 3, p. 716–727, 2013. Citado na página 25.


ICMC
USP
SÃO CARLOS