

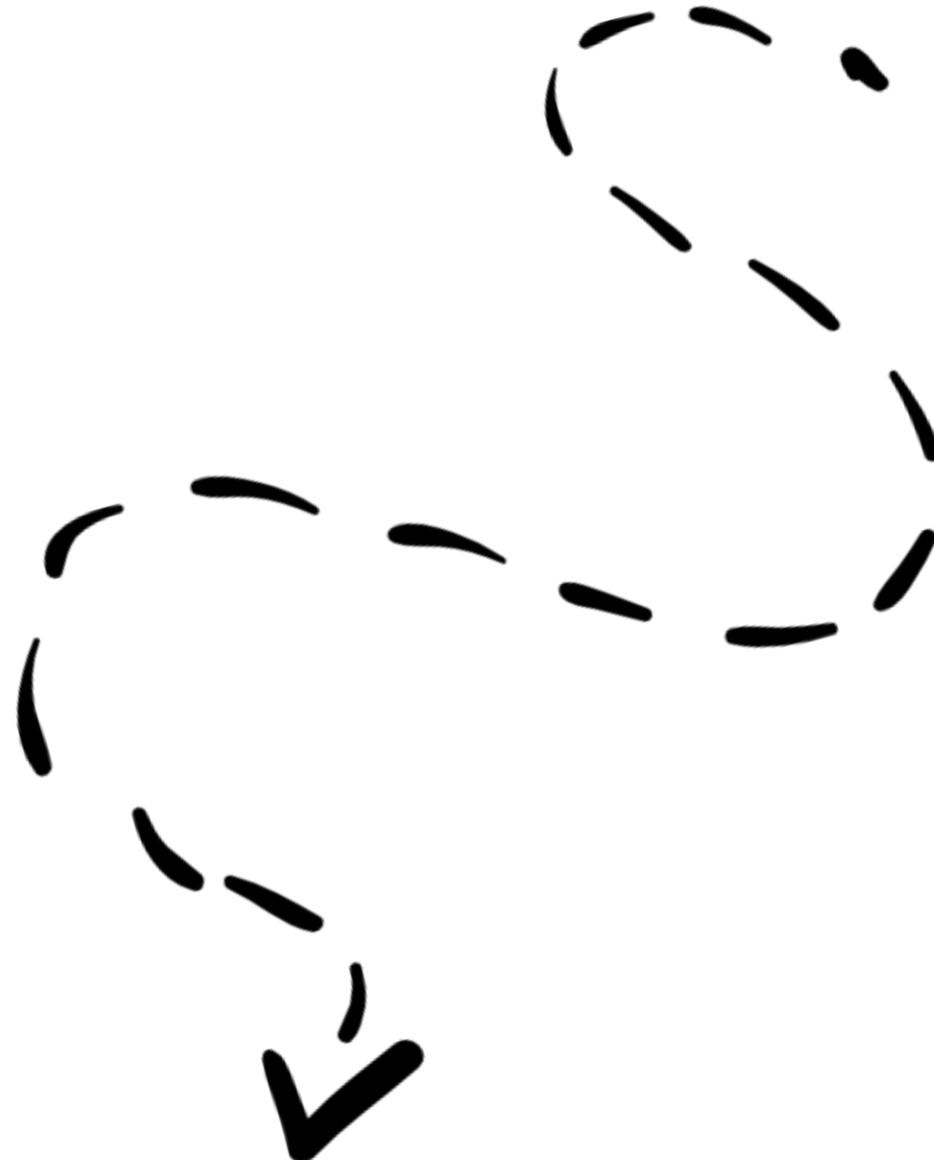


INTRODUÇÃO À ENGENHARIA E CIÊNCIA DE DADOS

Teoria e prática do começo ao
fim da história da informação

Agenda

- Apresentação geral
- Aula 0: Processo geral de dados
- Aula 1: Ambiente de execução
- Aula 2: Ingestão
- Aula 3: Preparação
- Aula 4: Enriquecimento
- Aula 5: Visualização
- Aula 6: Inteligência
- Avaliação



Conheça o instrutor



Miguel Sarraf F. Santucci

Formado em Engenharia de Computação pela Poli-USP na turma de 2021 e mestrando em Ciência da Computação pelo IME-USP. Líder Técnico na Stefanini Data & Analytics. 4 anos+ trabalhando com engenharia e ciência de dados para grande empresas do mercado. Diversas certificações na área de importantes provedores de nuvem e serviços especializados, além de outros muitos cursos e experiências.

[Conheça mais aqui.](#)



Ninguém sairá daqui especialista ou um engenheiro de dados completo. Vamos apresentar tópicos gerais da área e os principais métodos utilizados atualmente junto com um pouco de contexto de experiências passadas e boas práticas.

Vamos à alguns dados técnicos do curso:

- 10h totais de curso
- Conteúdos teóricos
- Apresentação de soluções exemplo desenvolvidas
- Proposta de desafio técnicos incremental ao longo das aulas
- 5 provas intermediárias pontuais
- 1 "Estudo de Caso" ao final
- 1 projeto opcional

Sistemas avançados

Aplicações de Dados

Banco de Dados

Fundamentos de programação

Programação Orientada à Objetos

Algoritmos e Estruturas de Dados

Programação básica

Software básico

Sistemas Operacionais

Compiladores

Programação de Sistemas

Dispositivos físicos

Arquitetura de Computadores

Eletrônica

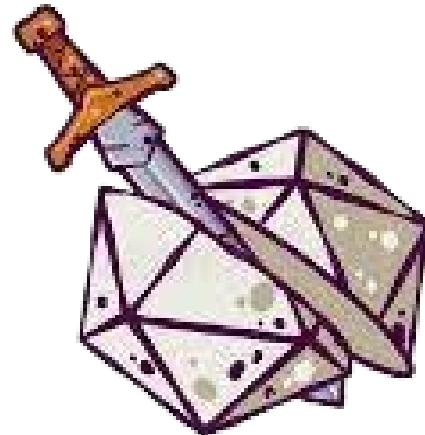
Circuitos Elétricos

Nós estamos aqui

Porque eles
vieram antes
de nós

AULA 0

PROCESSO GERAL DE DADOS



O que são boas práticas?

Cenário "Real"

As estruturas são confusas e adquirimos o conhecimento do domínio ao longo do tempo.

Complexo
Não sabemos o que não sabemos.

As boas práticas gerais podem precisar de ajustes

Cenário "Apocalíptico"

Não há estrutura e as regras não são bem definidas.

Caótico

Ein?

Boas práticas gerais não se aplicam.

Complicado

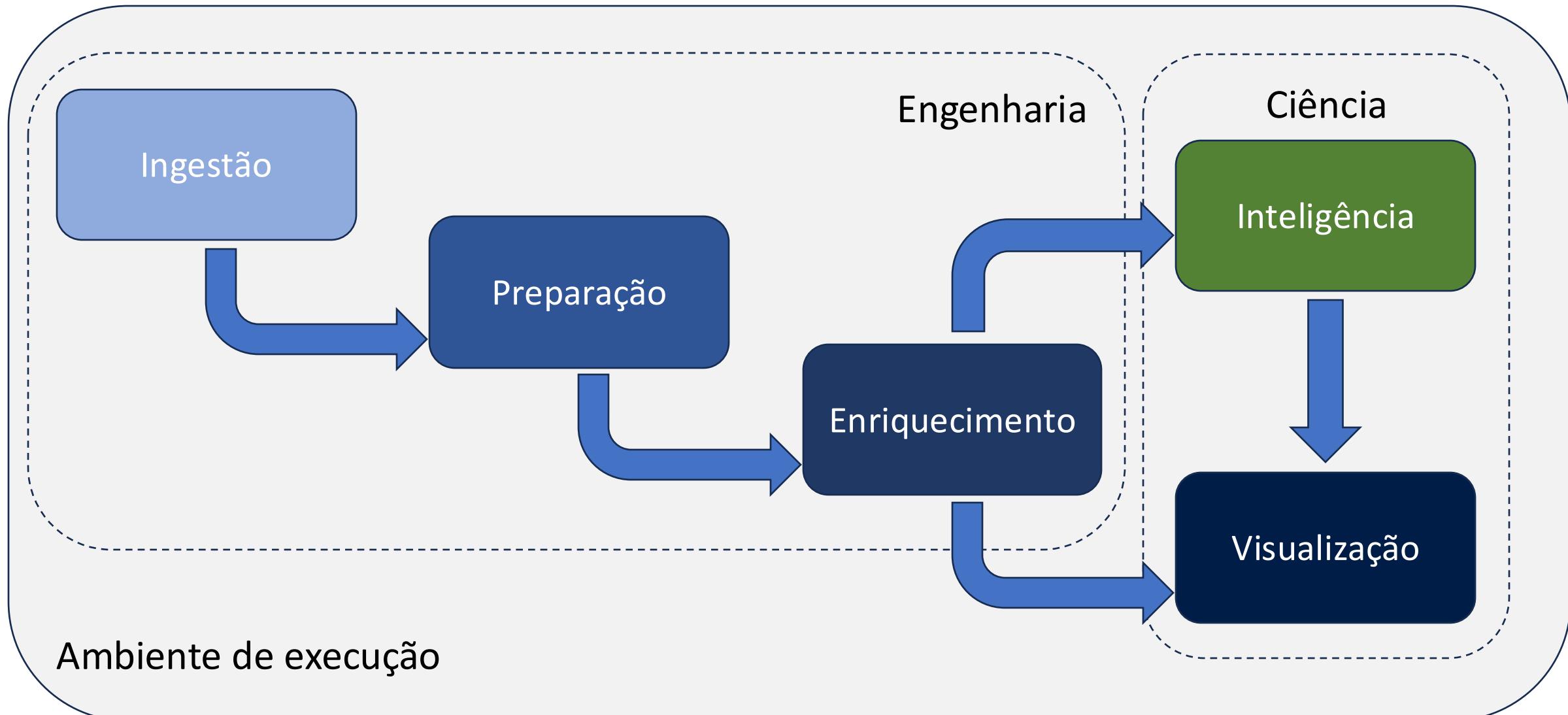
Sabemos o que não sabemos.

Cenário "Acadêmico"

As estruturas são previsíveis e temos todo o conhecimento do domínio trabalhado.

As boas práticas gerais se aplicam.

Processo geral de dados



ZONA BRONZE OU CRUA

Recebe os dados brutos tal qual chegam ao sistema

ZONA PRATA OU CURADA

Recebe os dados confiáveis: tratados, filtrados e completos

ZONA OURO OU ENRIQUECIDA

Recebe os dados valorizados, com cálculos avançados e regras de negócio

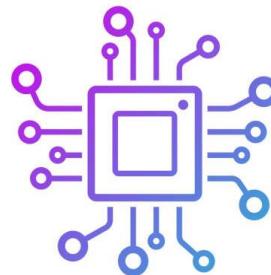


AULA 1

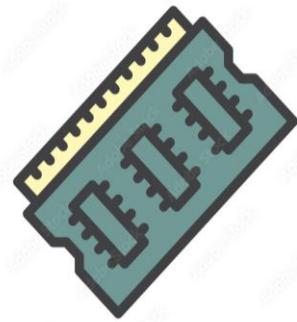
AMBIENTE DE EXECUÇÃO



A plataforma sobre a qual o processo desenvolvido será executado afeta diretamente o desempenho e as possibilidades. Surgem questões fundamentais:



...

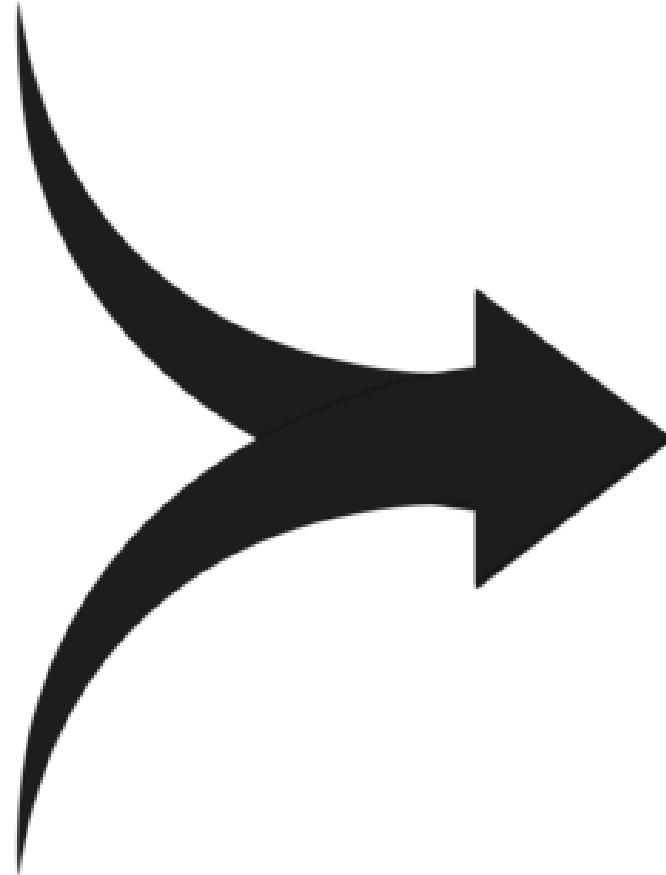


AULA 2

INGESTÃO



Ingestão - geral



Sistema em desenvolvimento



Cópia **idêntica** dos dados de origem.

Previne perda de informação em caso de problemas no processamento em camadas mais avançadas

Ambiente controlado

A ingestão por fluxo trata de trazer para dentro do sistema dados com atualização frequente ou que chegam de forma imprevisível.

Podem aparecer em qualquer formato que seja mais conveniente ao emissor.

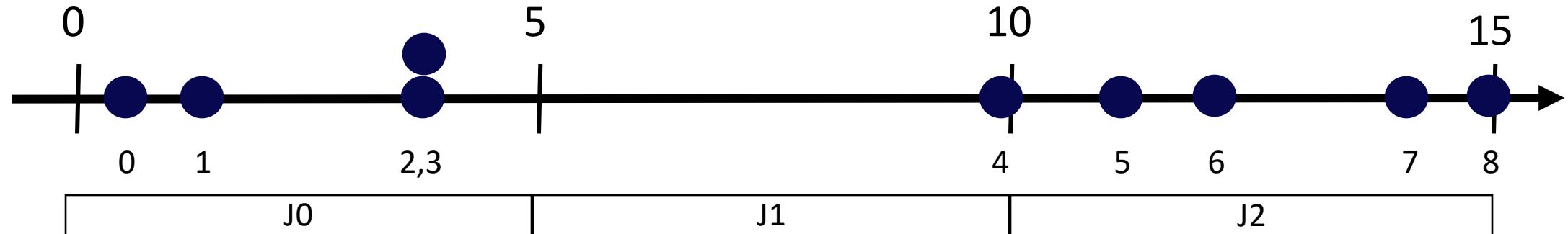
Como são em grande volume, não se pode esperar para tratar esses dados, mas também não se pode tratar todos assim que chegam.

A ingestão por lote trata de trazer para dentro do sistema dados com atualização infrequente na fonte.

No geral, são dados que chegam de forma estruturada e razoavelmente padronizados.

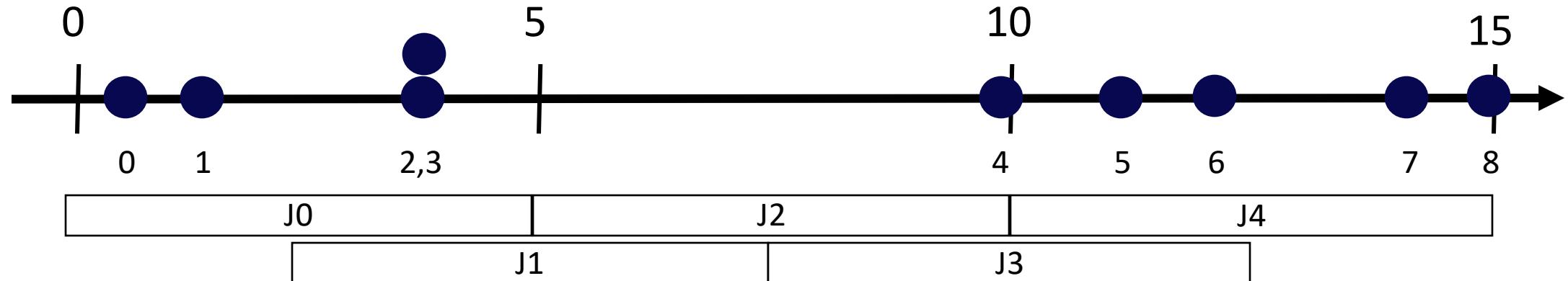
O tratamento desses dados pode ser agendado para horários convenientes sem perdas.

Janela em cascata

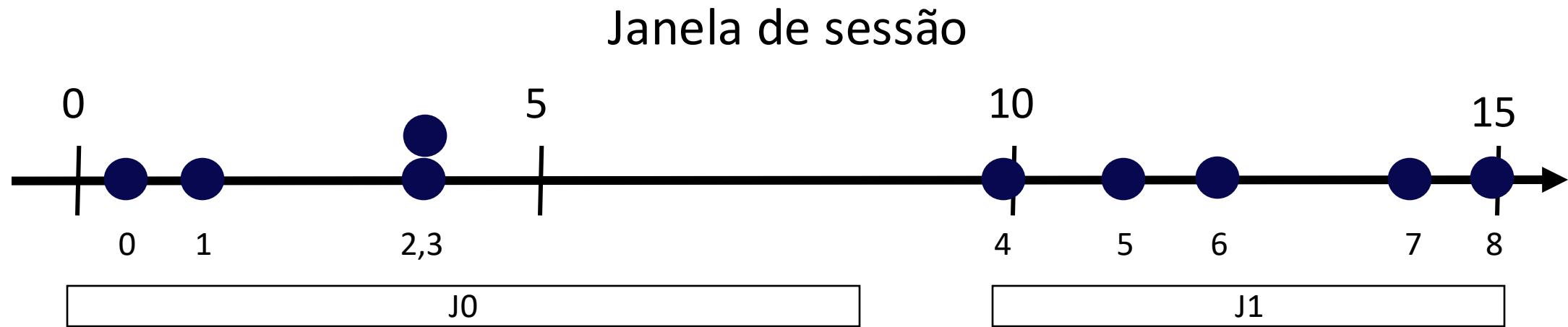


	E0	E1	E2	E3	E4	E5	E6	E7	E8
J0	X	X	X	X					
J1					X				
J2						X	X	X	X

Janela de salto

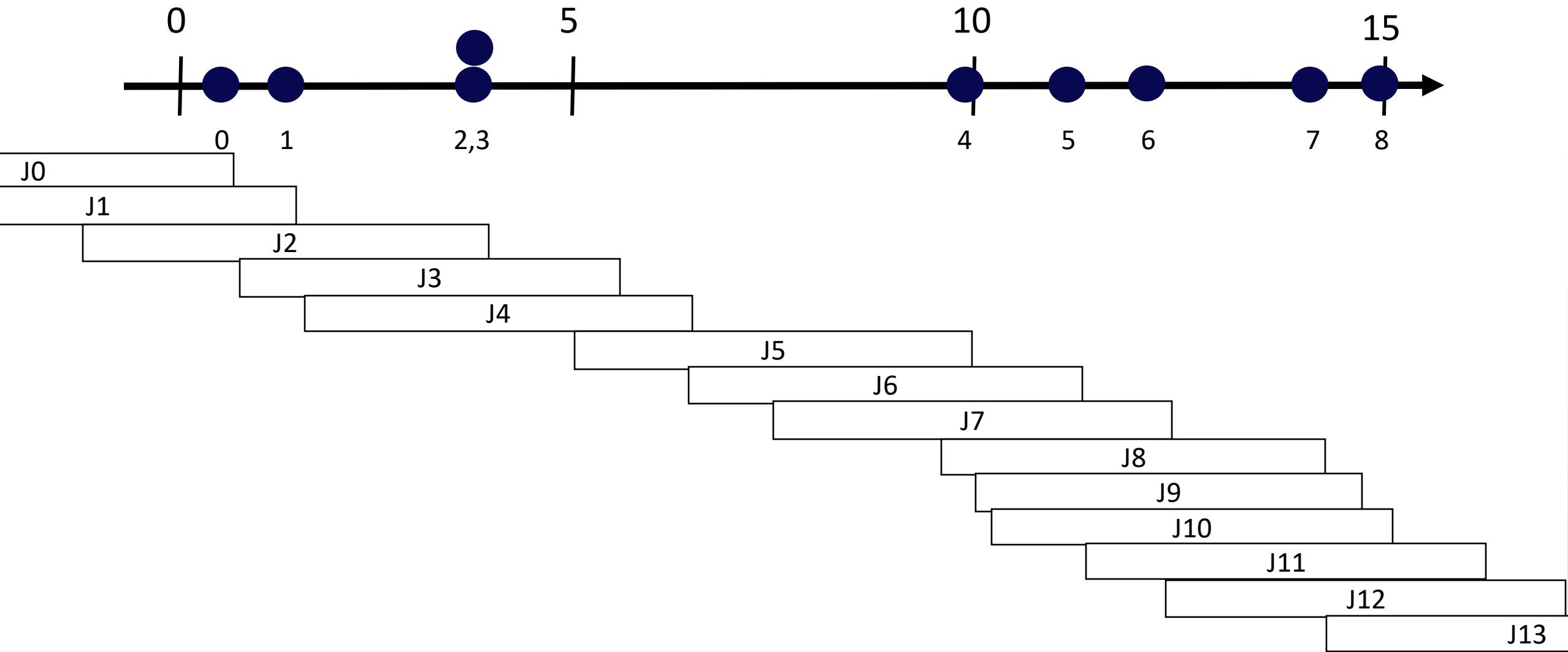


	E0	E1	E2	E3	E4	E5	E6	E7	E8
J0	X	X	X	X					
J1			X	X					
J2					X				
J3					X	X	X		
J4						X	X	X	X



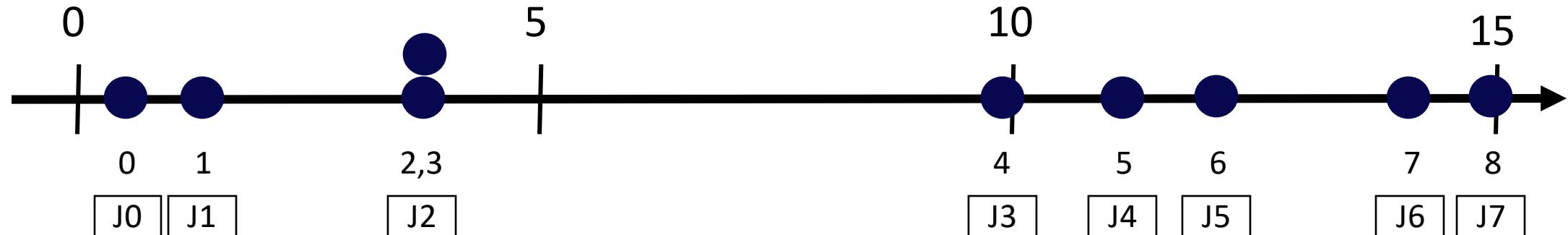
	E0	E1	E2	E3	E4	E5	E6	E7	E8
J0	X	X	X	X					
J1					X	X	X	X	X

Janela deslizante



Ingestão - Janelamento

Janela instantânea



	E0	E1	E2	E3	E4	E5	E6	E7	E8
J0	X								
J1		X							
J2			X	X					
J3					X				
J4						X			
J5							X		
J6								X	
J7									X

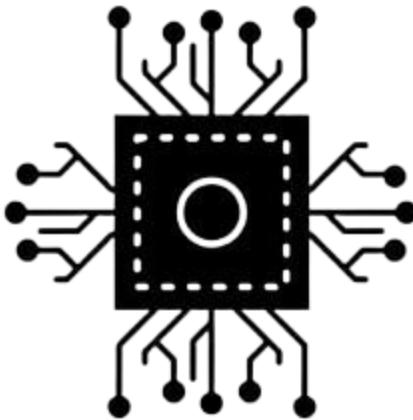
Ingestão - comparação de janelas



	Lote	Fluxo cascata	Fluxo Salto	Fluxo Sessão	Fluxo Deslizante	Fluxo Instantânea
Número de janelas	9	3	5	2	13	8
Total de eventos	9	9	14	9	35	9
Média de eventos	1	3	2.8	4.5	2.5	1.13
Variância dos eventos	0	3	1.7	0.5	1.19	0.13
Desvio Padrão dos eventos	0	1.73	1.3	0.71	1.09	0.35

Dados à obra!

Fontes



Monitor de informações da CPU.
Arquivos são escritos periodicamente em um diretório especificado.
Configure esse processo localmente!



Informações meteorológicas da sua cidade por hora.
Dados disponibilizados pelo INMET via download do site.
[Busque a sua cidade!](#)

Saídas

Arquivos tabelares contendo os dados de entrada, janelados ou não e indexados por data e hora.

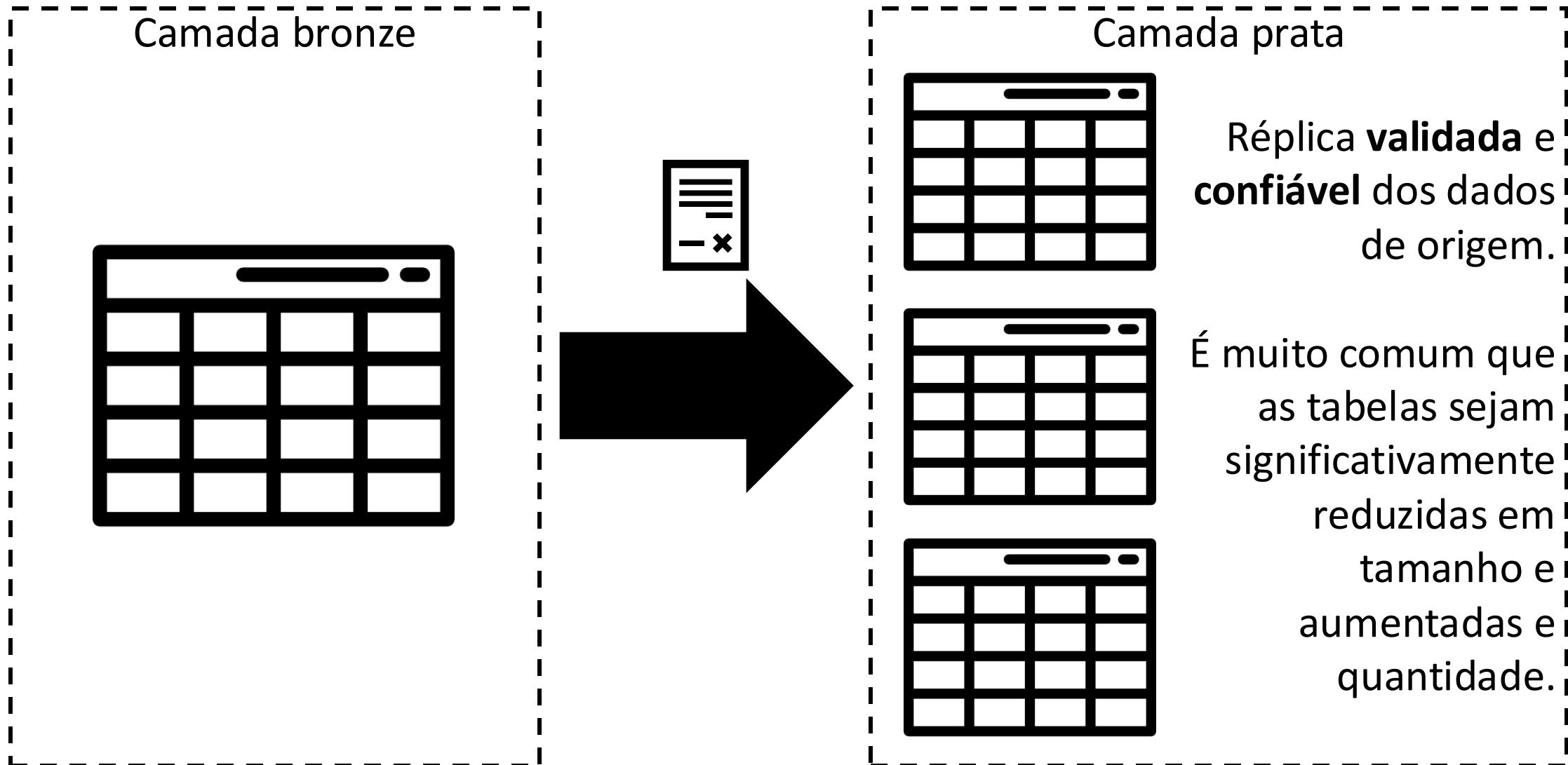
Defina formato do arquivo, forma de ingestão e estrutura de diretórios.

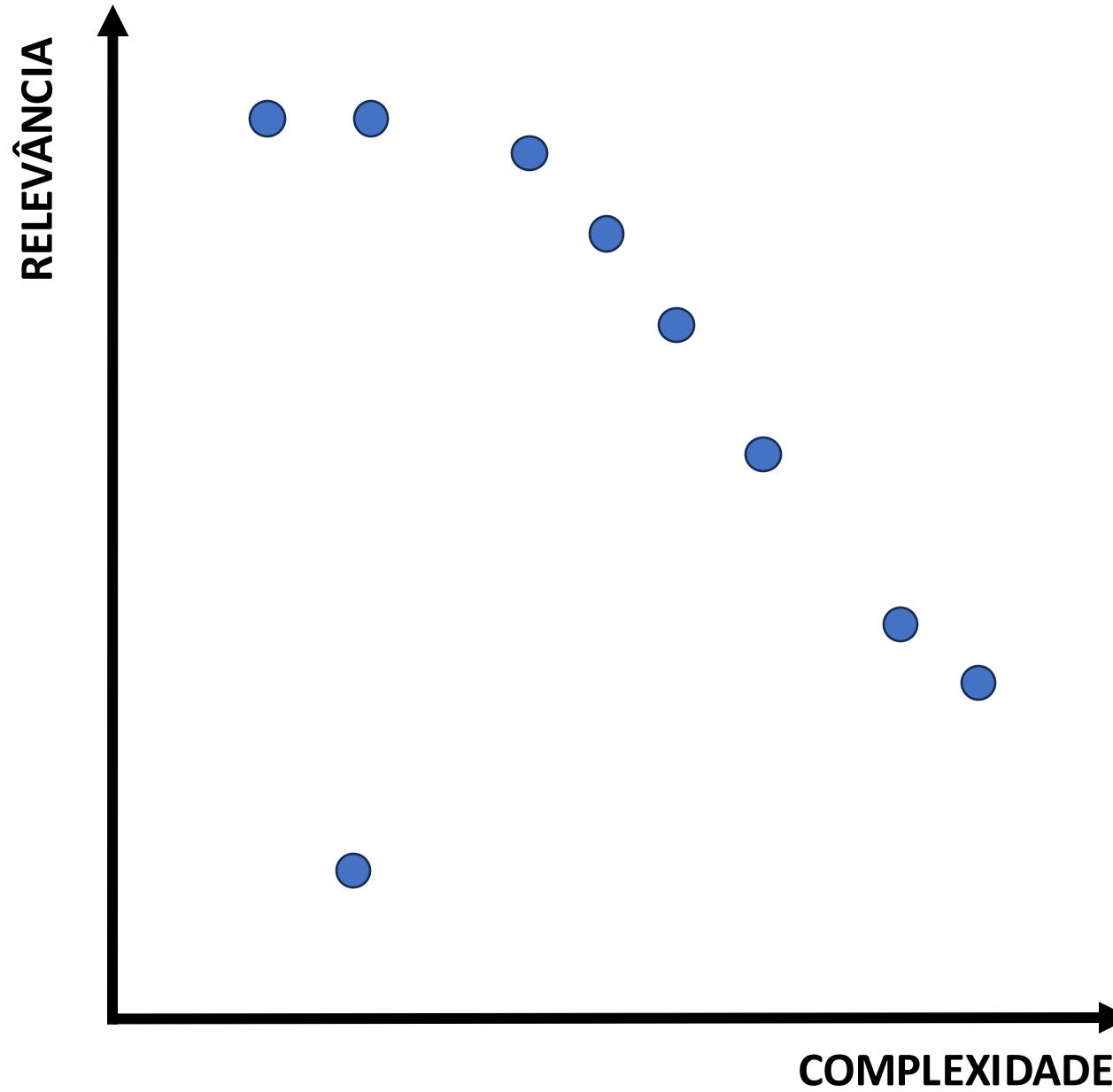


AULA 3

PREPARAÇÃO







Limpeza

Remoção de dados incompletos ou inconsistentes.

Formatação

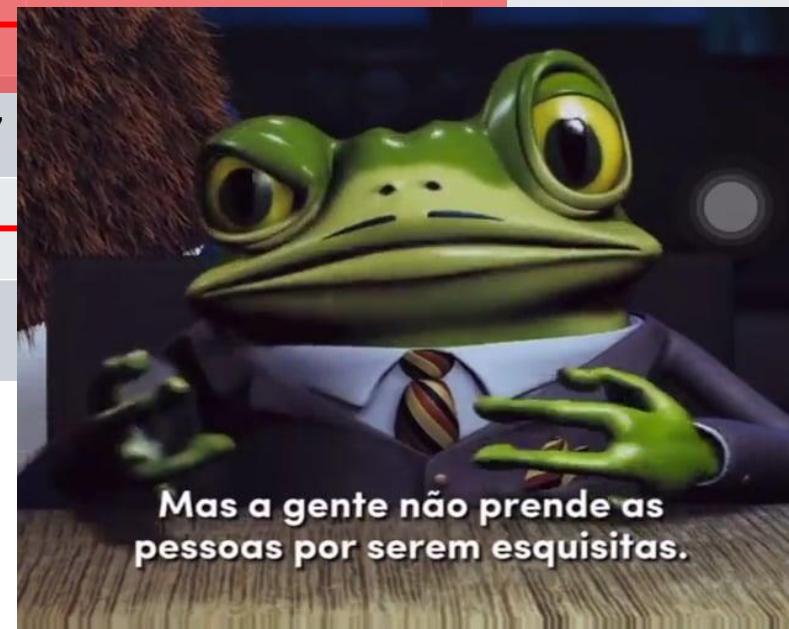
Ajuste do tipo de dado armazenado em cada coluna, com atenção especial nos limites impostos.

Normalização

Correção das tabelas para que se tornem mais estruturadas e compactas, ajudando na separação lógica dos dados e evitando anomalias. Existem algumas Formas Normais (FN) a serem aplicadas incrementalmente.

Preparação - limpeza

empresa	abertura	cnpj	num_funcionarios	valor	ativo
ABC	17/10/1999	51.563.537/0001-09	30	R\$10.000,00	1
XYZ	31/08/1991			R\$20.500,00	0
UMDOISTRES	18/02/1997			R\$12.345,67	1
	25/06/2001			R\$99.999,99	1
XPTO	10/05/1991			R\$10,00	0



Mas a gente não prende as pessoas por serem esquisitas.

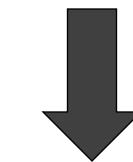
Dados faltantes

Dados inválidos

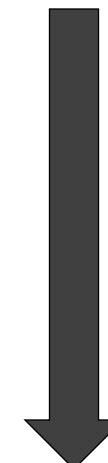
de
inverossímeis

Preparação - formatação

empresa	abertura	cnpj	num_funcionarios	valor	ativo
ABC	17/10/1999	51.563.537/0001-09	30	R\$10.000,00	1
UMDOISTRES	18/02/1997	14.601.825/0001-76	10	R\$12.345,67	1
XPTO	10/05/1991	93.375.016/0001-60	20	R\$10,00	0



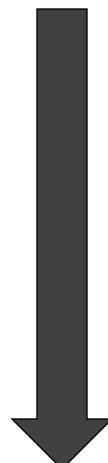
string/texto



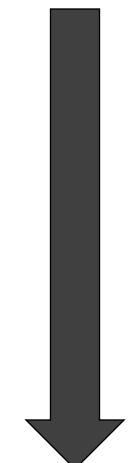
string/texto
Embora composto
apenas de números, é
melhor tratar esse tipo
de dado como texto.

data

Necessário considerar diferentes
formatos de data e padrões.



float/numérico
Pode ser necessário
remover máscaras,
considerar diferentes
marcadores decimais e
de milhares.



booleano

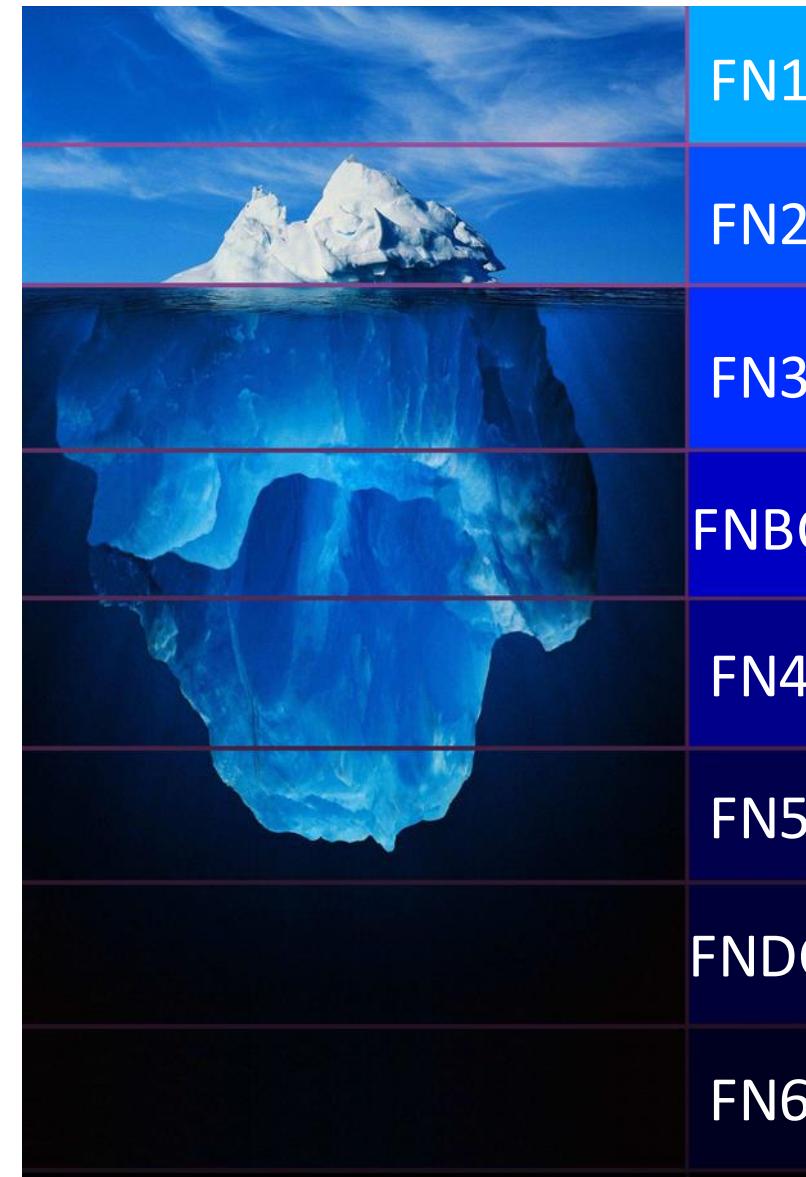
Conversão de tipo
pode ser necessária

Conversão de tipo pode
ser útil em alguns casos.

Preparação - normalização

Anomalias de banco de dados são comportamentos que fogem ao esperado durante a utilização de um SGBD.

Nesse processo, o número de tabelas sempre aumenta e, colateralmente, o tamanho delas é reduzido.



Atomicidade

Dependência parcial

Dependência transitiva

Dependência funcional

Dependência multivalorada

Dependência de junção

Restrições mínimas

Irreducibilidade

Anomalias funcionais

Anomalias de redundância

Casos especiais

Atomicidade: Os elementos de cada campo devem ser indivisíveis

nome	nascimento	idade	endereco	contato
Joaquim	17/10/1999	25	Rua dos Alfeneiros, 4	jmanuel@gmail.com
Ana	18/02/1997	28	Av. X, 9	(11) 0234-5678
Pedrita	10/05/1991	34	Rua Y, 2000, apto 101	pe.drita@hotmail.com thelittlerock@gmail.com
Enzo	04/05/2010	15	Estr. da Vitoria, 23, apto 4	(33) 9748-1745 (22) 1607-5874
Valentina	28/12/2006	18	Wallaby Way, 42	anitnelav@uol.com.br (56) 1545-6043

- Logradouro
- Número
- Complemento

- Email 1
- Email 2
- Telefone 1
- Telefone 2

Dependência parcial: Atributos dependem unicamente da chave primária

aluno	disciplina	professor	creditos
Joaquim	Bancos de Dados	Beogival Dante	3
Ana	Introdução à Python	Ivete Pereira	2
aluno	disciplina	professor	creditos
Joaquim	Bancos de Dados	Beogival Dante	3
Ana	Introdução à Python	Ivete Pereira	2
Pedrita	Engenharia de Software	Eleonor Santos	2
Joaquim	Engenharia de Software	Eleonor Santos	2
Ana	Bancos de Dados	Alice Curá	3
Pedrita	Arquitetura de Computadores		

Dependência transitiva: Atributos não chave devem ser independentes

disciplina	professor	creditos	tipo_vinculo
Bancos de Dados	Beogival Dante	3	Prof. convidado
Introdução à Python	Ivete Pereira	2	Prof. doutor
Engenharia de Software	Eleonor Santos	2	Prof. Pós-doc

disciplina	professor	creditos	professor	tipo_vinculo
Bancos de Dados	Beogival Dante	3	Beogival Dante	Prof. convidado
Introdução à Python	Ivete Pereira	2	Ivete Pereira	Prof. doutor
Engenharia de Software	Eleonor Santos	2	Eleonor Santos	Prof. Pós-doc
Arquitetura de Computadores	Alice Curá	3	Alice Curá	Prof. doutor

Dependência funcional: os atributos de cada tabela devem ter relação direta com a sua função

sala	dia_horario	eh_prof	tipo_reserva	tipo_sala	eh_prof	tipo_reserva
Sala 404	14/07/2025 14:00:00	1	aula	sala	1	aula
Laboratório de redes	01/01/2024 06:00:00	1	experimento	laboratório	1	experimento
		0		sala	0	monitoria
Sala 42	04/05/2026 16:20:00	0		laboratório	0	sessão de estudo

Dependência multivalorada: Permutação de chaves categóricas deve ser representada em mais de uma tabela

biblioteca	livro	curso
Bib. Luis Olegário Brandão	Banco de Dados – Projeto e Implementação	ADS
Bib. Luis Olegário Brandão	Banco de Dados – Projeto e Implementação	SI
Bib. Luis Olegário Brandão	Entendendo Algoritmos	ADS
Bib. Luis Olegário Brandão	Entendendo Algoritmos	SI

biblioteca	livro	atado	biblioteca	curso
Bib. Luis Olegário Brandão	Banco de Dados – Projeto e Implementação	itado	Bib. Luis Olegário Brandão	ADS
Bib. Luis Olegário Brandão	Entendendo Algoritmos	itado	Bib. Luis Olegário Brandão	SI
Bib. Alan Mathison Turing	Redes de Computadores	de Co	Bib. Alan Mathison Turing	SI
Bib. Alan Mathison Turing	Organização Estruturada de Computadores	de Co	Bib. Alan Mathison Turing	Eng. Da Comp.

Dependência de junção: É possível separar a tabela em diversas outras e retornar à tabela original

fornecedor	componente
Mamute	resistor
Mamute	capacitor
FilipeFlop	osciloscópio
FilipeFlop	capacitor
Baú da Eletrônica	FPGA

laboratorio	fornecedor	componente
Lab. Eletrônica	Mamute	resistor
Lab. Eletrônica	Mamute	capacitor
Lab. Eletrônica	Mamute	osciloscópio
Lab. Eletrônica	FilipeFlop	resistor
Lab. Sistemas digitais	Mamute	capacitor
Lab. Sistemas digitais	FilipeFlop	osciloscópio
Lab. Sistemas digitais	Baú da Eletrônica	resistor
Lab. Sistemas digitais	Baú da Eletrônica	capacitor
Lab. Sistemas digitais	Baú da Eletrônica	FPGA

Restrições mínimas: Relacionamento entre um atributo e faixas de um outro devem ser mantidas em tabelas separadas.

disciplina	nota	resultado	minimo	maximo
Redes de Computadores	5.5	Aprovado	5.0	10.0
Cálculo Numérico	4.3	Recuperação	3.0	4.9
Engenharia de Software	8.7	Aprovado	0.0	2.9
Arquitetura de Computadores	2.0	Reprovado		
Segurança da Informação	7.9			

Irredutibilidade: todas as tabelas são reduzidas para o formato chave-valor

nome	nascimento	idade	logradouro	numero	complemento
Joaquim	17/10/1999	25	Rua dos Alfeneiros	4	
Ana	18/02/1997	28	Av. X	9	

ID	nome	ID	nascimento	ID	idade	ID	lograd.	ID	ID	numero	ID	compl.
0	Joaquim	0	17/10/1999	0	25	0	Rua dos Alfeneiros	2	0	4	2	Apto 101
1	Ana	1	18/02/1997	1	28	1	Av. X	4	1	9	3	Apto 4
2	Pedrita	2	10/05/1991	2	34	2	Rua Y	2	2	2000		
3	Enzo	3	04/05/2010	3	15	3	Estr. da Vitoria	3	3	23		
4	Valentina	4	28/12/2006	4	18	4	Wallaby Way	4	4	42		



Fontes

Os arquivos resultado do processo de ingestão.

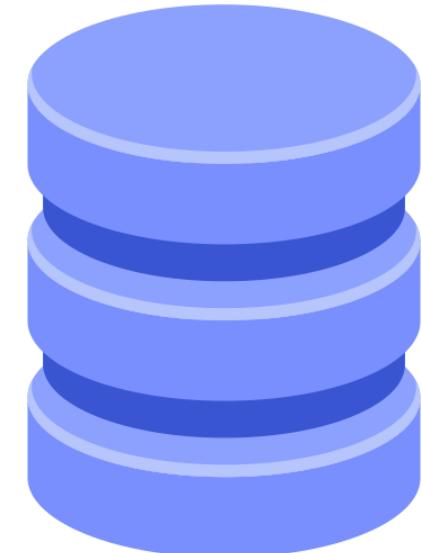
- Um arquivo contendo as medições dos parâmetros físicos da sua máquina pessoal;
- Um arquivo contendo o histórico meteorológico da sua cidade em um período concomitante de tempo

Saídas

As mesmas informações, porém limpas, formatadas e normalizadas e salvas em um banco de dados.

Será preciso estudar o esquema dessas tabelas de entrada para definir os passos necessários e configurar a base de dados corretamente.

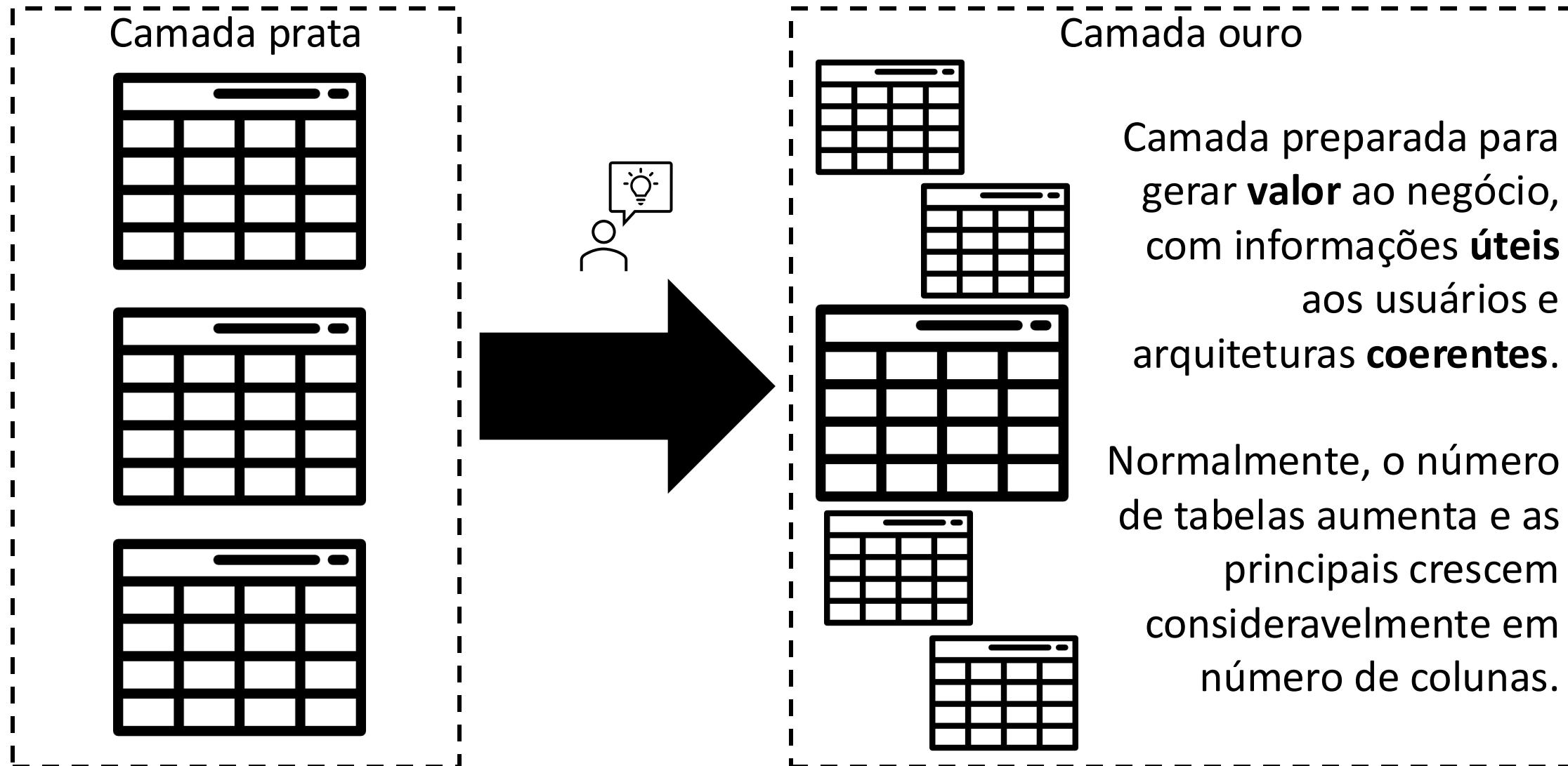
Naturalmente, poderão existir mais de uma tabela de saída por tabela de entrada.



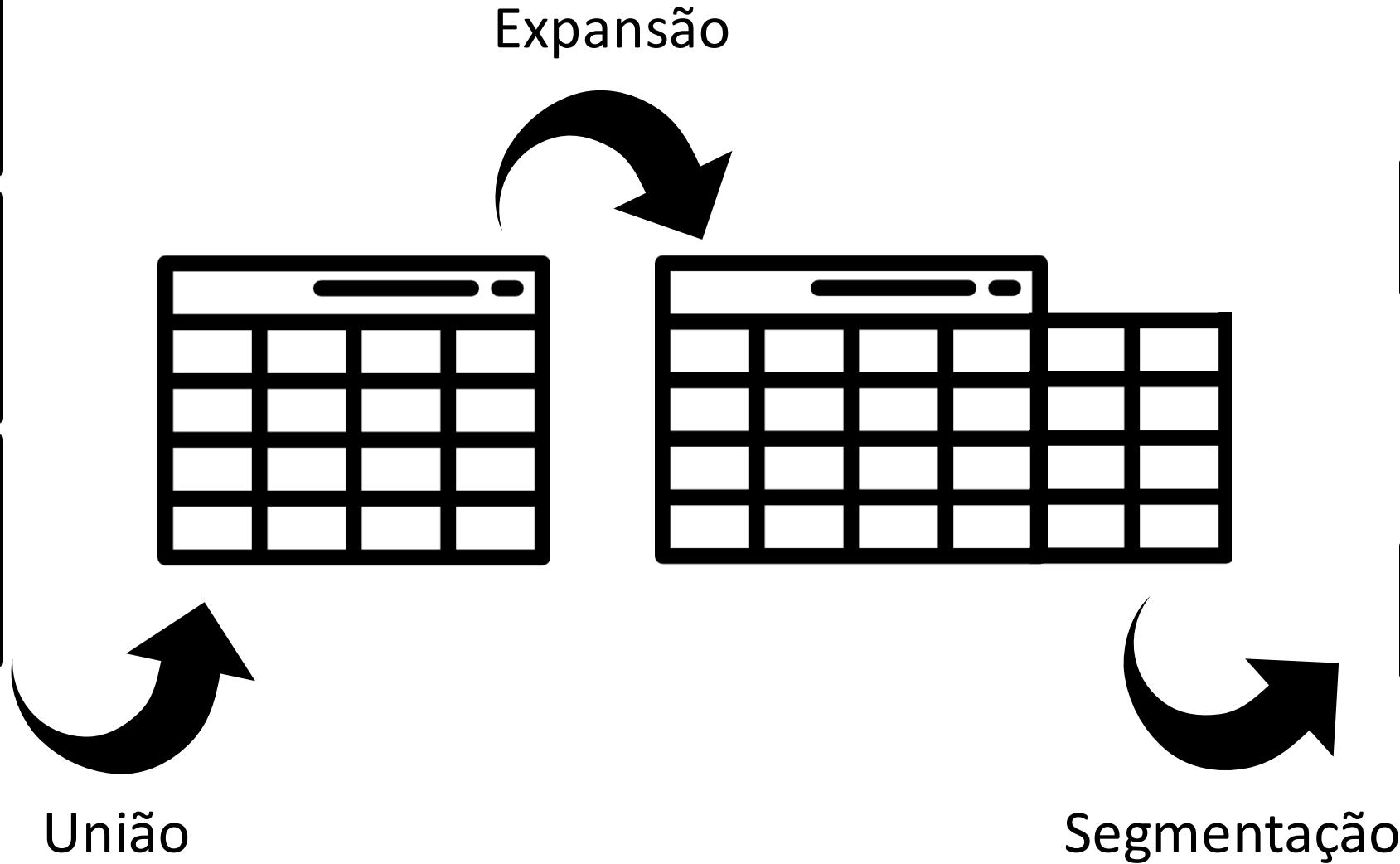
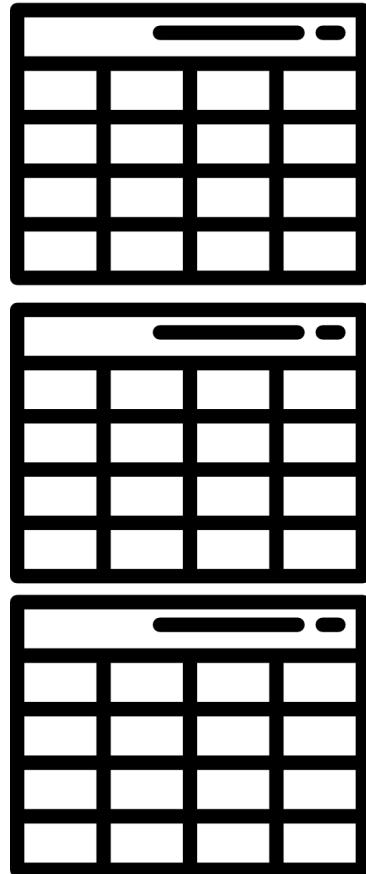
AULA 4

ENRIQUECIMENTO





Enriquecimento - métodos



Enriquecimento - união

Tabelas com conteúdos correlatos devem ser unidas para que todo o domínio dessa informação esteja centralizado.

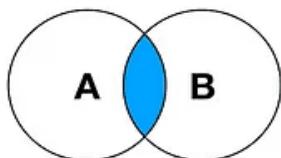
municipio	populacao	pib	idh
Belo Horizonte	2.315.560	41.818,32	0.810
Caxambu	21.056	17.204.61	0.743
Ibirité	170.537	17.407,52	0.704
Conselheiro Lafaiete	131.621	23.881,55	0.761

municipio	mercado_A	mercado_B	mercado_C
Belo Horizonte	800.000	700.000	500.000
Caxambu	3.000	12.000	5.000
Diamantina	15.000	15.000	15.000
Ibirité	10.000	20.000	100.000

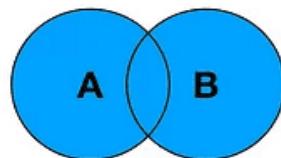


municipio	populacao	pib	idh	mercado_A	mercado_B	mercado_C
-----------	-----------	-----	-----	-----------	-----------	-----------

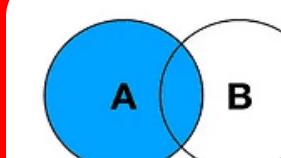
Enriquecimento - união



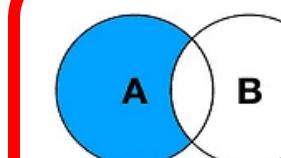
INNER JOIN



FULL OUTER JOIN



LEFT JOIN



LEFT JOIN EXCLUDING
INNER JOIN

municipio	populacao	pib	idh	mercado_A	mercado_B	mercado_C
-----------	-----------	-----	-----	-----------	-----------	-----------

municipio	populacao	pib	idh
Belo Horizonte	2.315.560	41.818,32	0.810
Caxambu	21.056	17.204,61	0.743
Ibirité	170.537	17.407,52	0.704
Conselheiro Lafaiete	131.621	23.881,55	0.761

municipio	mercado_A	mercado_B	mercado_C
Belo Horizonte	800.000	700.000	500.000
Caxambu	3.000	12.000	5.000
Diamantina	15.000	15.000	15.000
Ibirité	10.000	20.000	100.000

Enriquecimento - expansão

Novas métricas de negócio são criadas para enriquecer a análise dos usuários e minimizar processamento na próxima camada. Conhecimento do negócio e bom senso são imprescindíveis para essa etapa.

municipio	populacao	pib	idh	mkt_shr_A	mkt_shr_B	mkt_shr_C	disponivel
Belo Horizonte	2.315.560	41.818,32	0,810	0,3454	0,3023	0,2159	0,1364
Caxambu	21.056	17.204,61	0,743	0,1425	0,5699	0,2375	0,501
Ibirité	170.537	17.407,52	0,704	0,0586	0,1173	0,5864	0,2377
Conselheiro Lafaiete	131.621	23.881,55	0,761				
Diamantina	45.000	←		0,3333	0,3333	0,3333	0

"Mini preparação":
Preencher valor nulo resultante da
etapa de junção

Aplicar de conhecimento do domínio
para criar novas informações
relevantes.

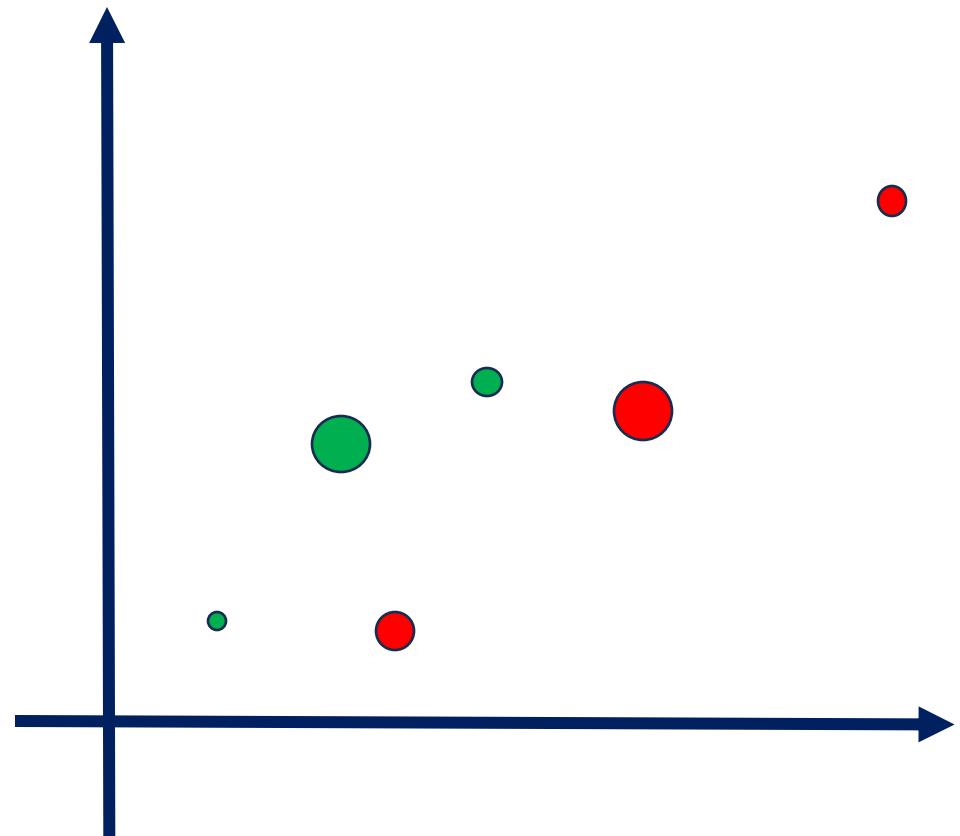
Nesta última etapa, a preocupação é em deixar todos os dados arrumados para a etapa de visualização. É preciso atentar à forma como cada dado vai ser utilizado para minimizar o processamento da camada seguinte. Para isso, é preciso conhecer os dois tipos essenciais de tabelas:

Dimensões

São as informações que delimitam os registros e serão usadas para separá-los e filtrá-los. Não trazem informações ao negócio.

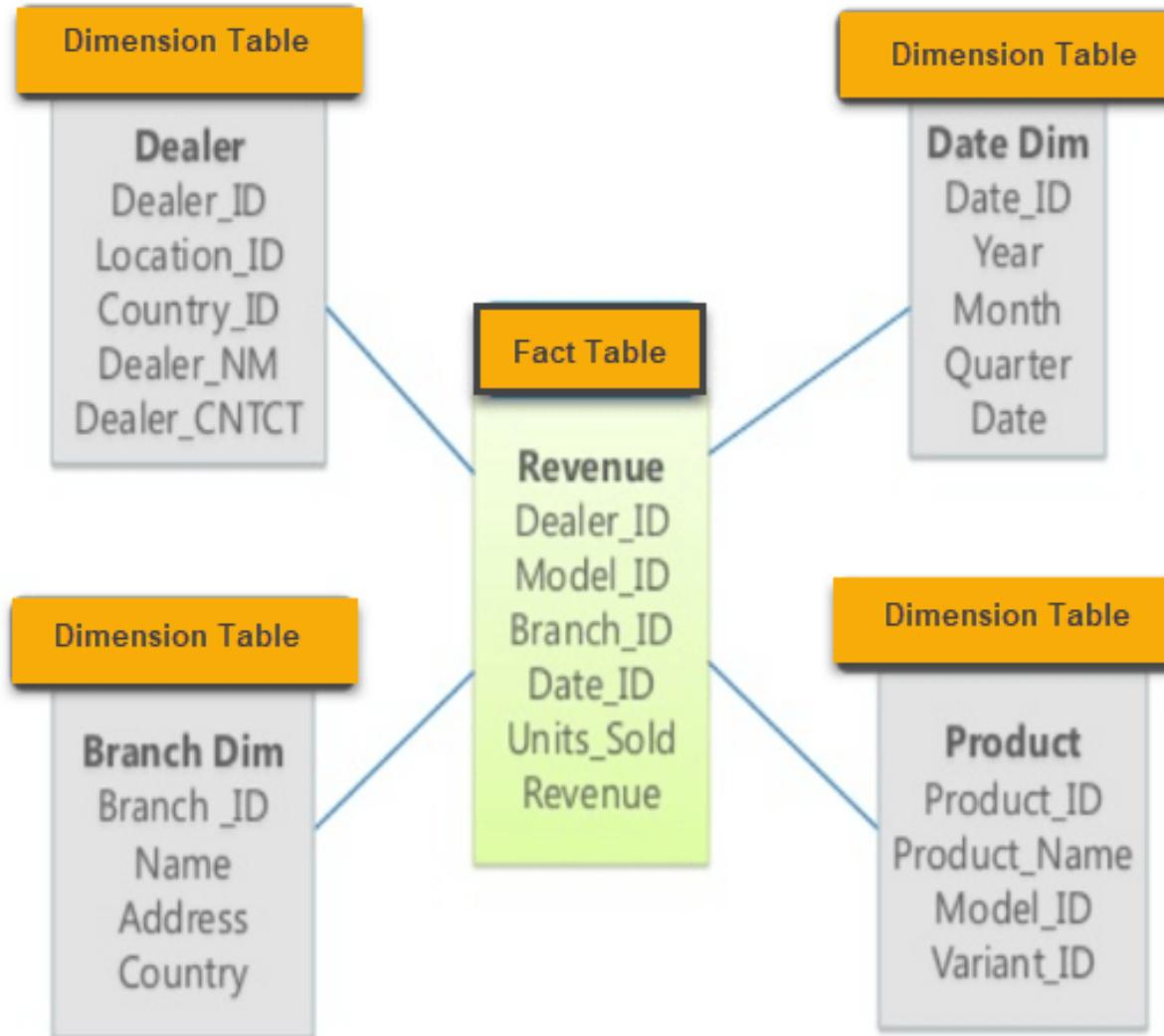
Fatos

São as informações realmente relevantes para o negócio, que explicam e são atribuídas aos registros para transmitir suas características.



Enriquecimento - segmentação

Enriquecimento - esquema estrela



O mais simples possível!

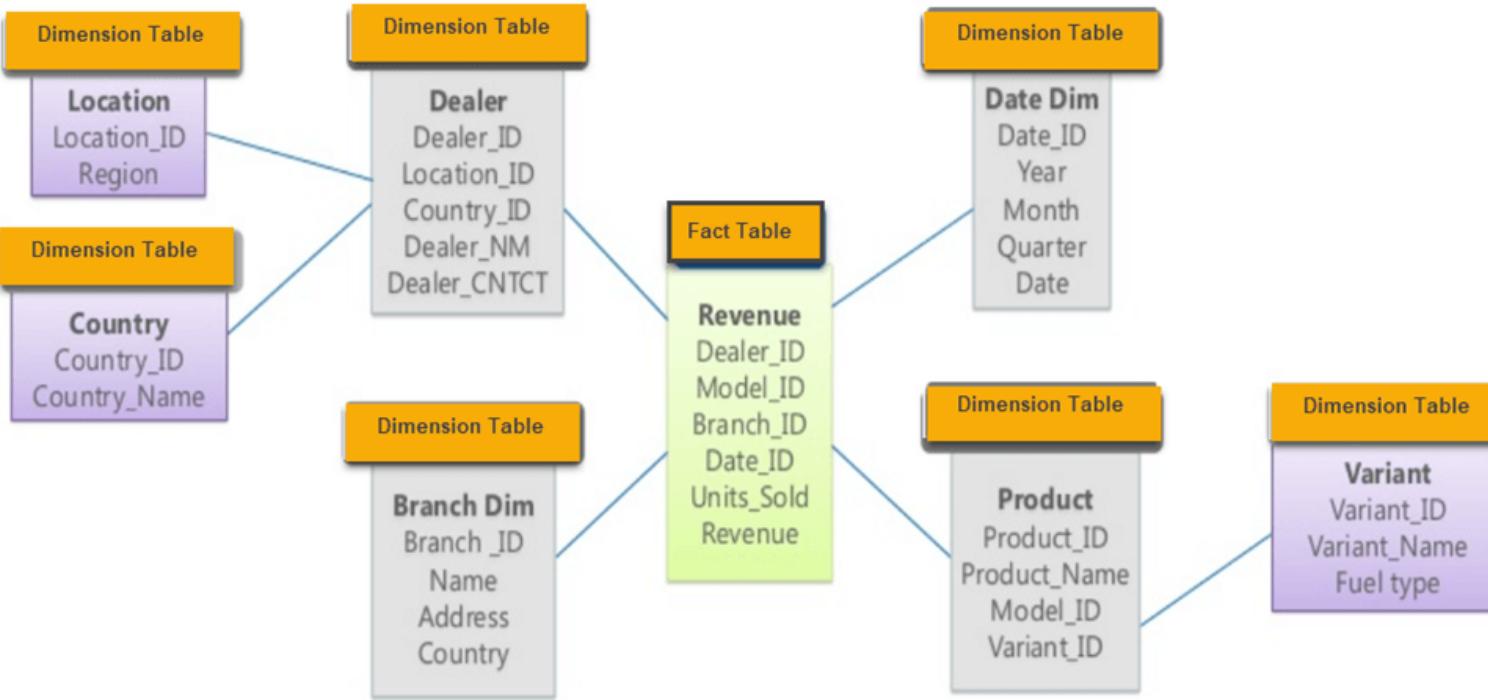
Consiste em uma tabela fato combinada com diversas tabelas dimensão.

Permite filtragens básicas e espelha um caso de negócio razoavelmente limitado.

É o mais eficiente e normalmente é possível reduzir outros esquemas a ele.

Nome alternativo (cunhagem própria):
Esquema de planeta

Enriquecimento - esquema floco de neve



As coisas começam a ficar interessantes!

Consiste em uma tabela fato combinada com diversas tabelas dimensão e cada dimensão tem suas próprias dimensões.

Permite filtragens avançadas e espelha um caso de negócio mais maduro.

Eficiência começa a ser penalizada, mas ainda viável.

Nome alternativo (cunhagem própria):
Esquema de sistema solar

Enriquecimento - esquema galáxia

A vida como ela é!



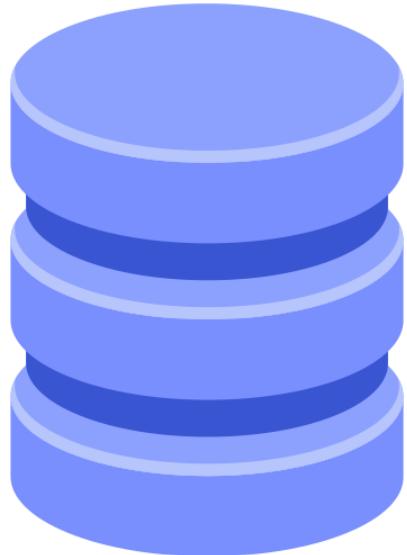
Em mais de uma tabela
interligadas por
dimensão e
suas
plexas e
negócio
to.

E necessário cuidado com a
eficiência no uso dos filtros.

Nome alternativo (cunhagem própria):
Esquema de Tatooine

Dados à obra!

Fontes



As tabelas resultado da camada anterior, com os dados criados organizados e confiáveis para tratamento.

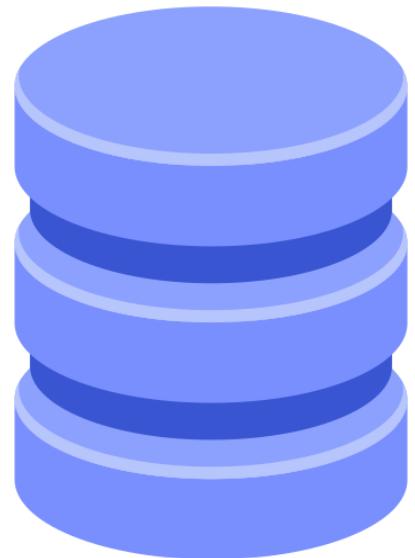
Cada uma das tabelas deve ser lida e entendida dentro do seu contexto de criação e seus dados revisados e ponderados.

Saídas

Agora é a hora de soltar a criatividade.

Devemos juntar e processar as tabelas conforme as relações entre os dados e colunas delas.

Busque entender o contexto e adquirir conhecimentos sobre possíveis relações entre as variáveis e como criar novos valores relevantes.

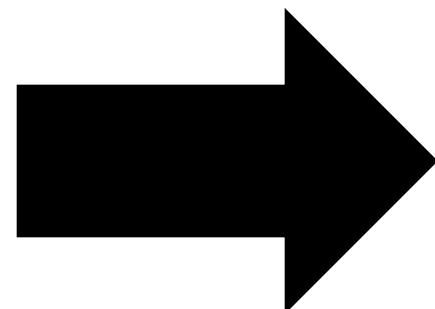
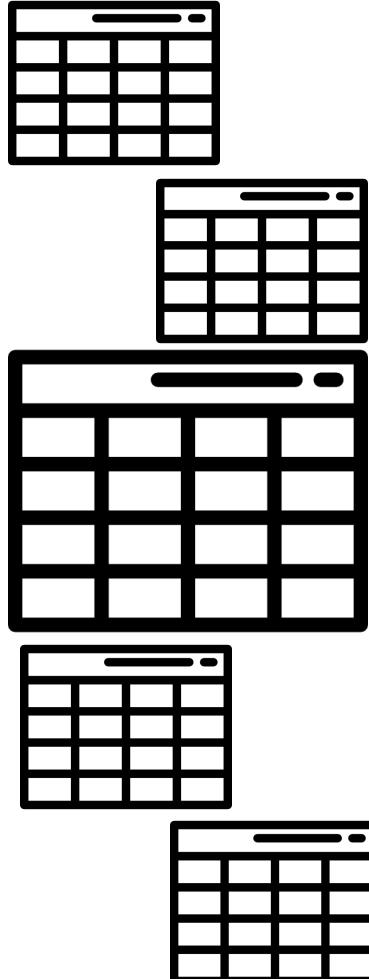


AULA 5

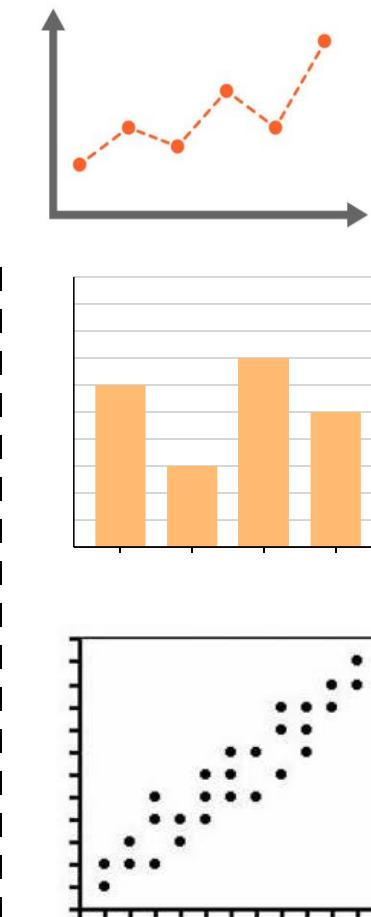
VISUALIZAÇÃO



Camada ouro



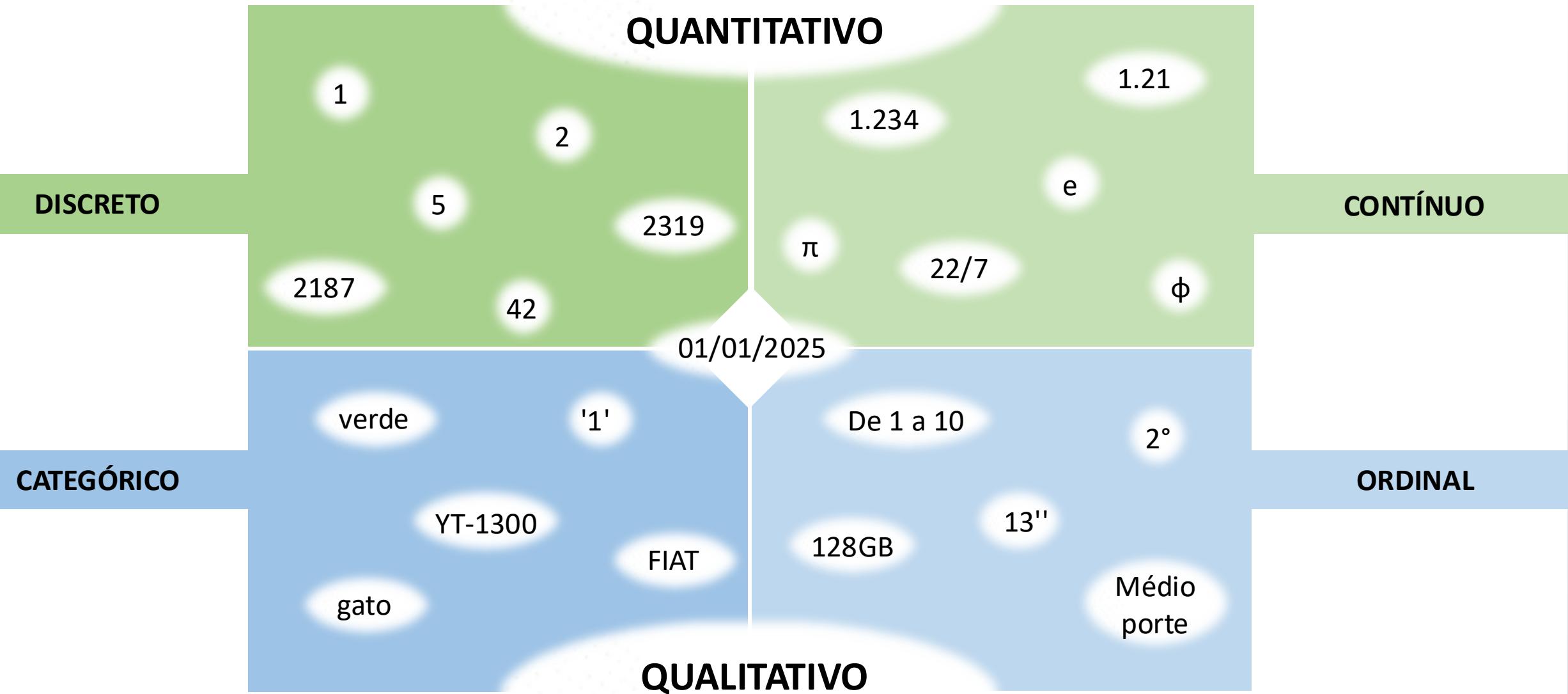
Relatórios de negócio



Os relatórios de negócio são os produtos finais consumíveis pelos usuários do sistema.

É através deles que serão extraídas as informações contidas nos dados e por onde será possível mostrar o valor agregado no processo desenvolvido.

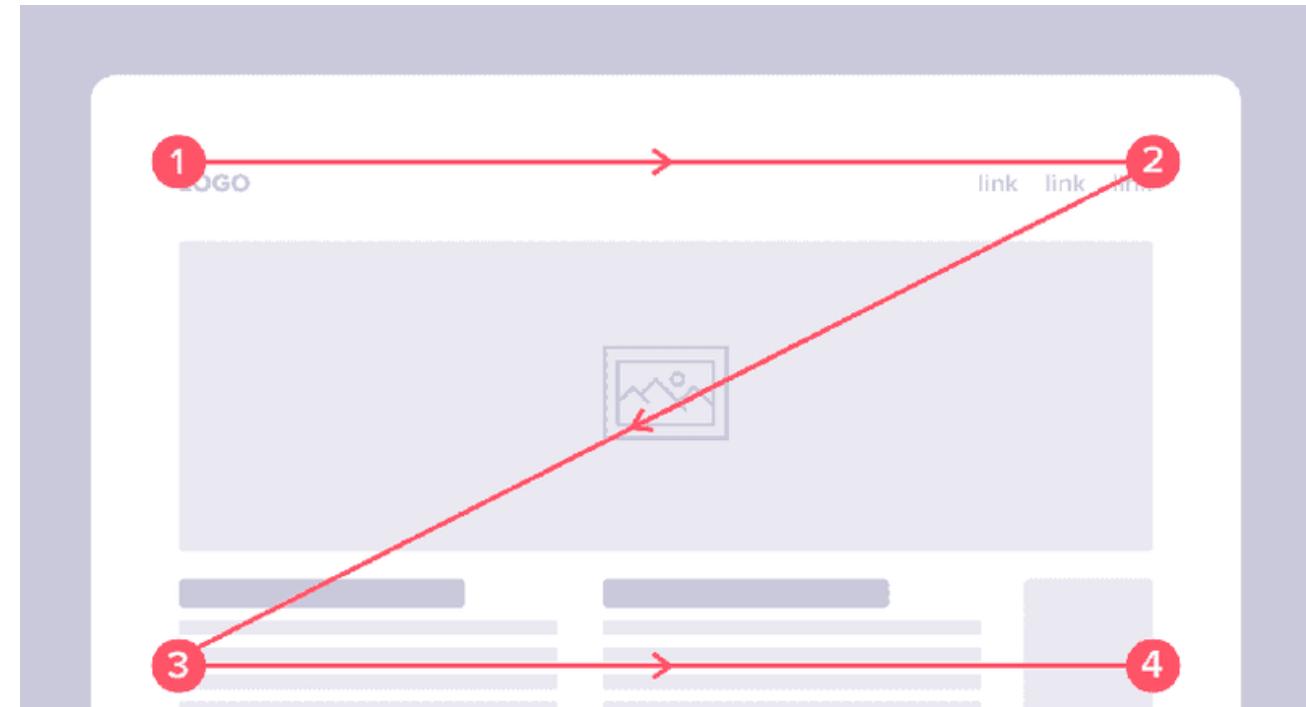
Visualização - tipos de dados

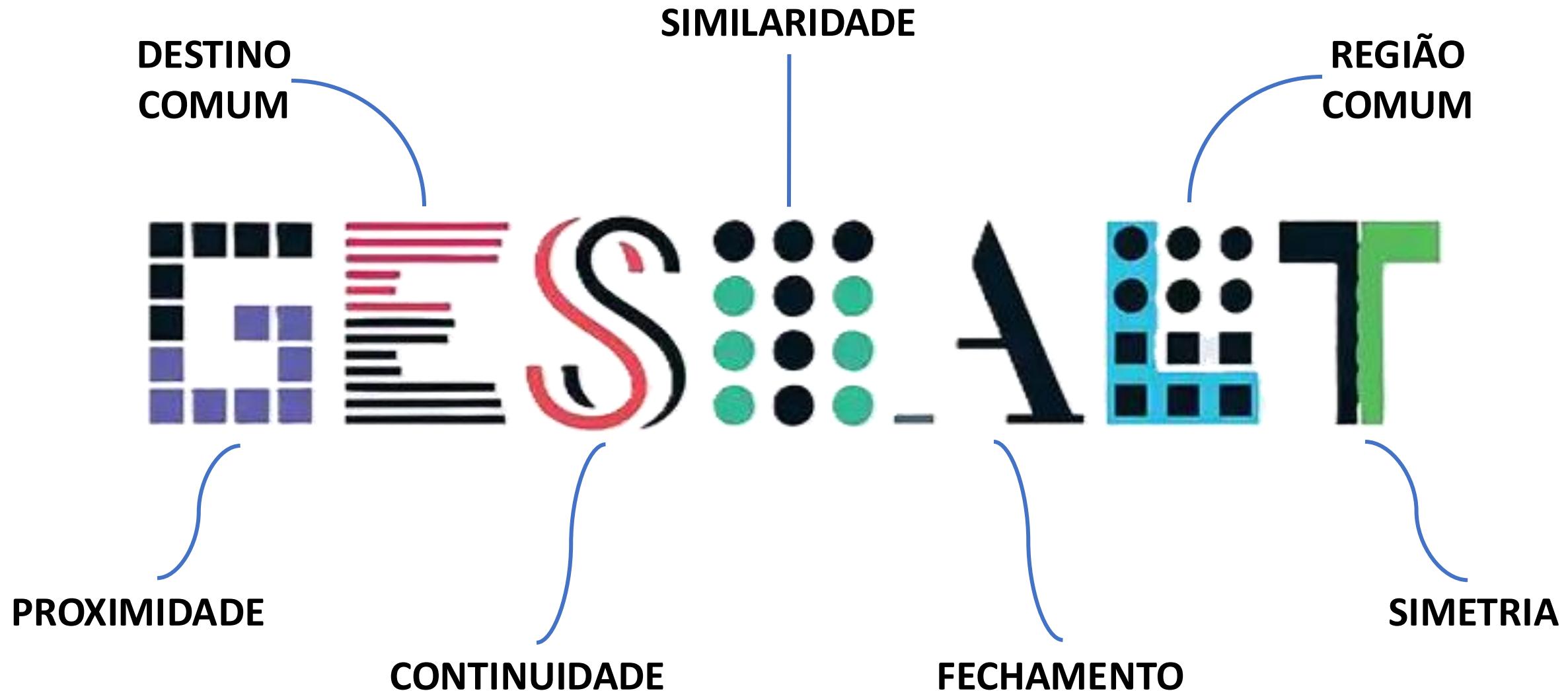


VOCÊ e você pode voltar para ler essa frase ou não.
VAI LER ISSO PRIMEIRO.
E depois vai ler isso em seguida.

Você vai querer ler este texto se desejar obter mais informações sobre o assunto, pois ele contém uma quantidade considerável de informações mesmo sendo compacto e com um espaçamento não ideal. Muitas pessoas provavelmente vão pular este parágrafo, por isso é importante dar atenção à hierarquia da informação.

E provavelmente você vai ler isso antes do parágrafo.



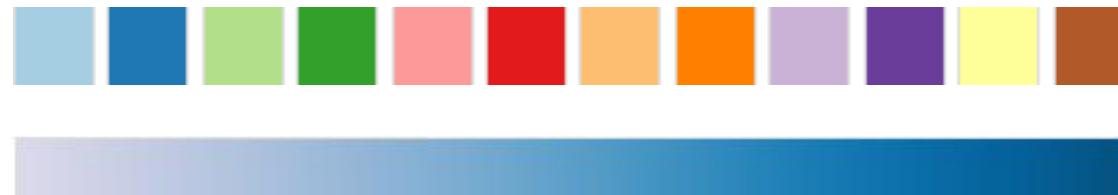


Visualização - cores

Conformidade
com palheta
predefinida



Boas práticas
gerais



Valores categóricos

Valores ordinais

Inclusão!!!



Visualização - tipos de gráficos

TIPOS DE DADOS

NARRATIVA



PRINCÍPIOS DE VISUALIZAÇÃO

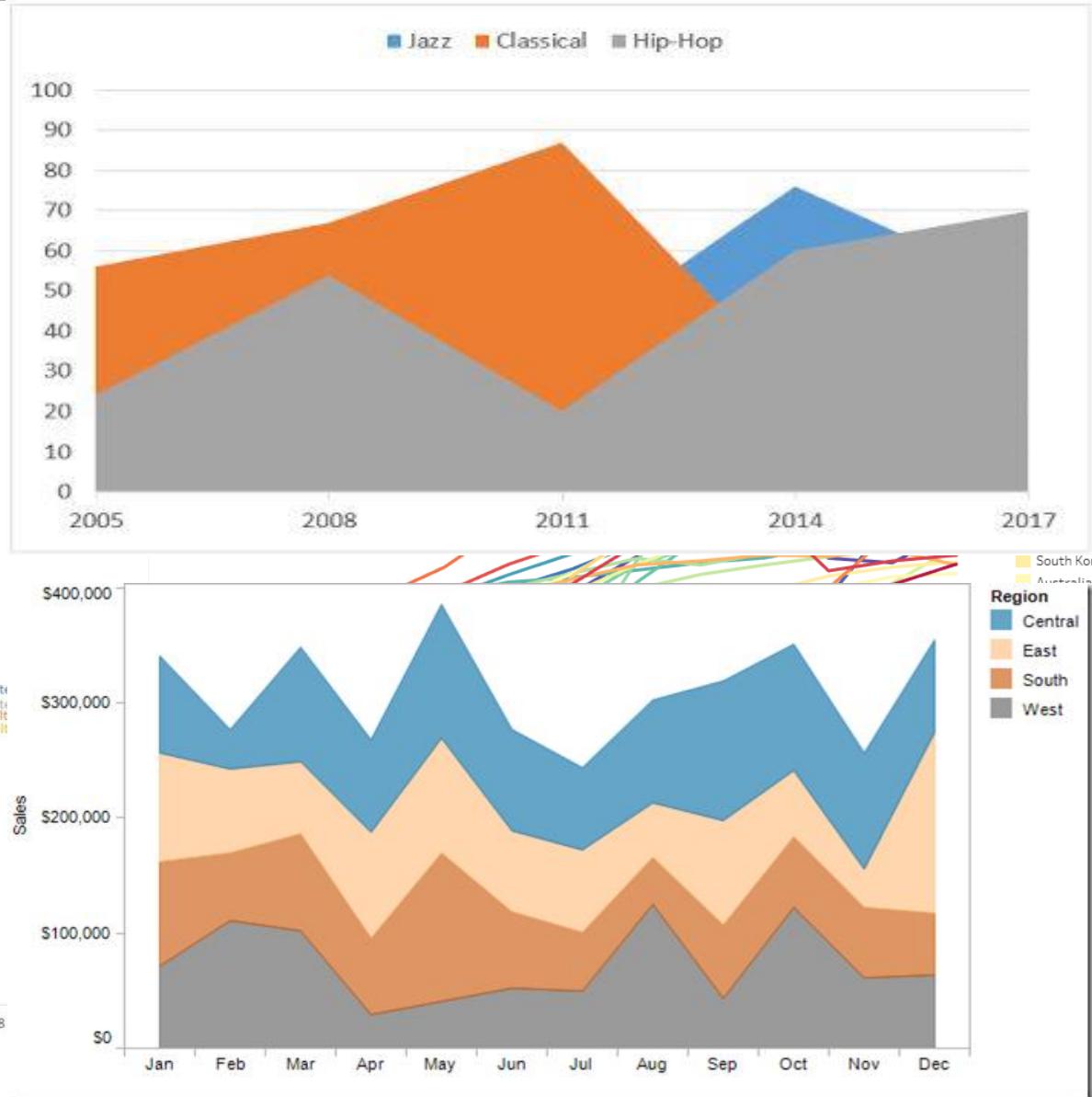
CORES

Visualização - gráficos de linha

Valores numéricos ao longo do tempo

A se pensar:

- Número de classes
- Linha ou área?

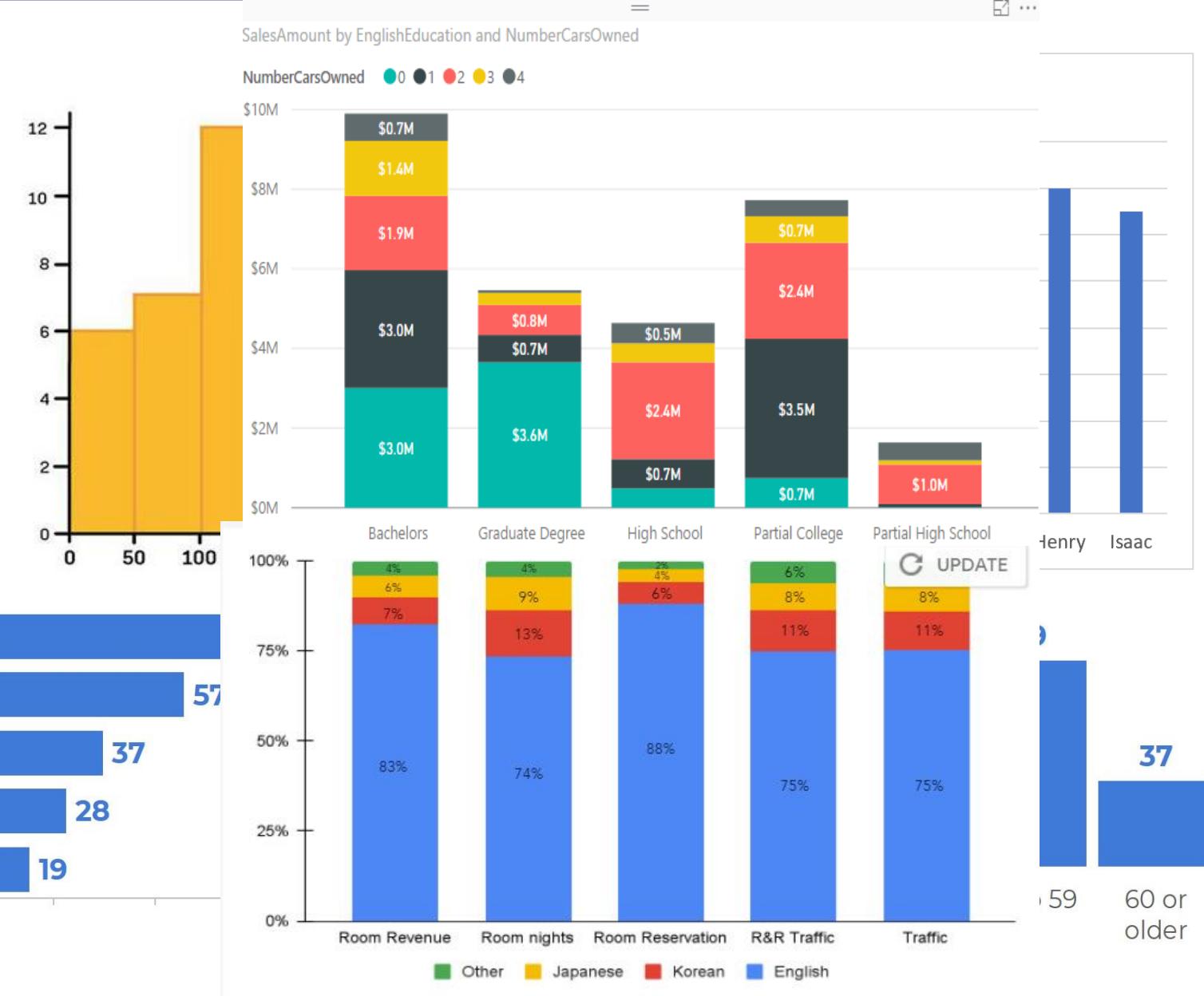
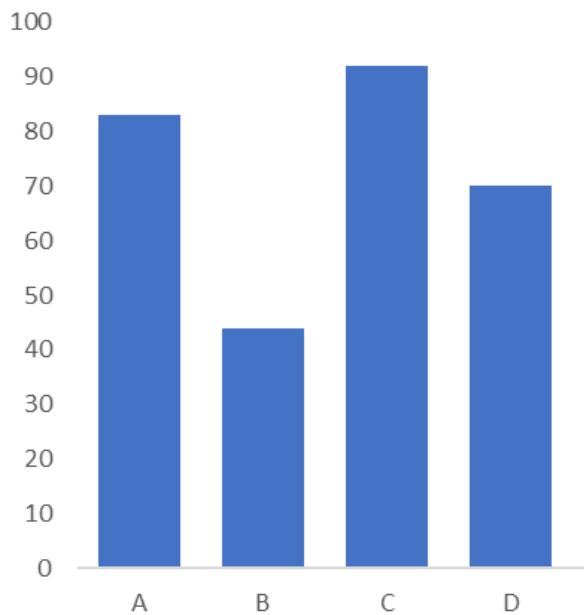


Visualização - gráficos de barras

Valores numéricos por classes

A se pensar:

- Largura das colunas
- Vertical ou horizontal?
- Simples ou empilhado?

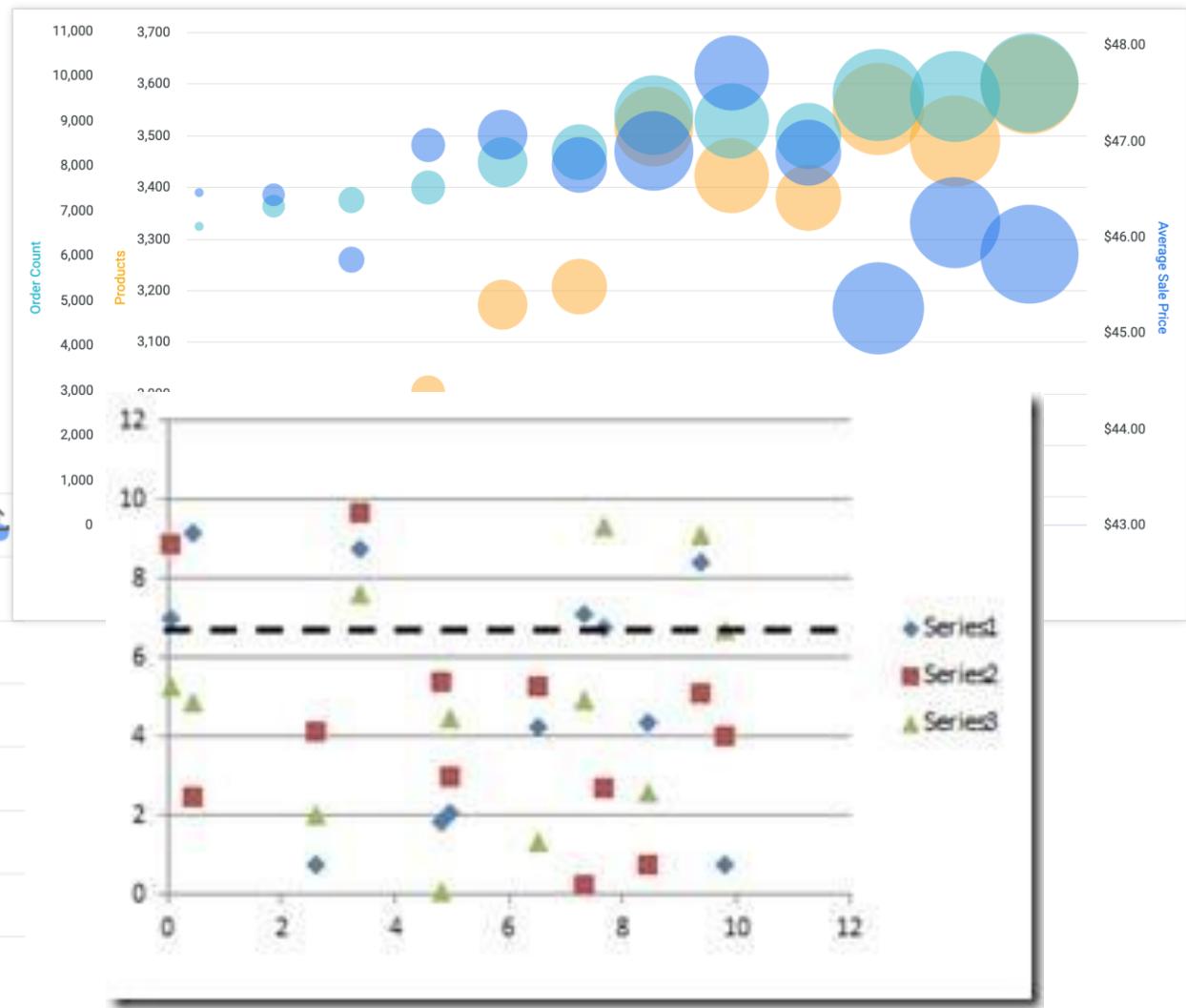
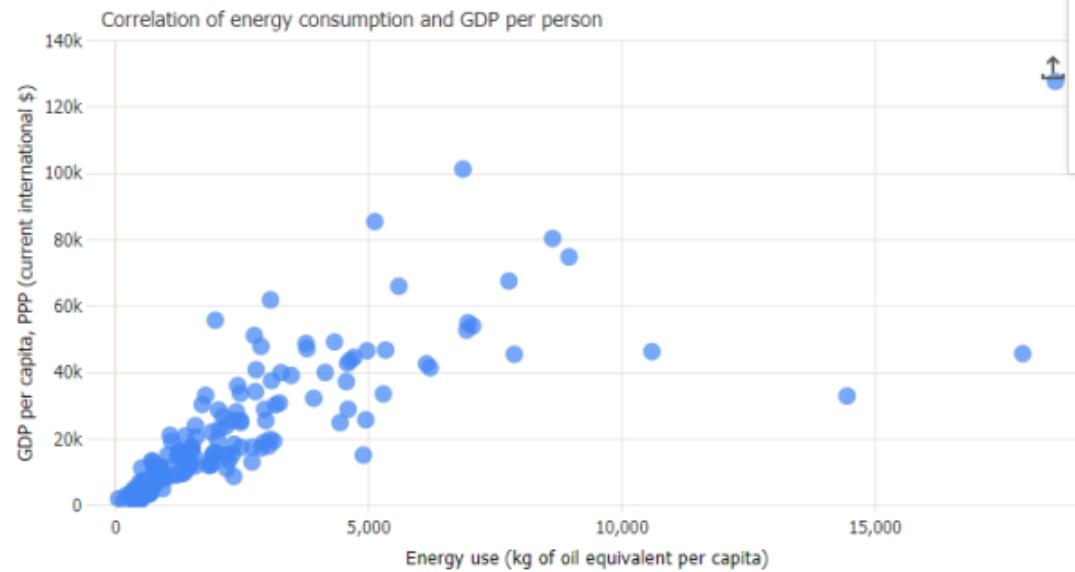


Visualização - gráficos de dispersão

Interação entre valores numéricos

A se pensar:

- Tamanho dos pontos
 - Formato dos marcadores

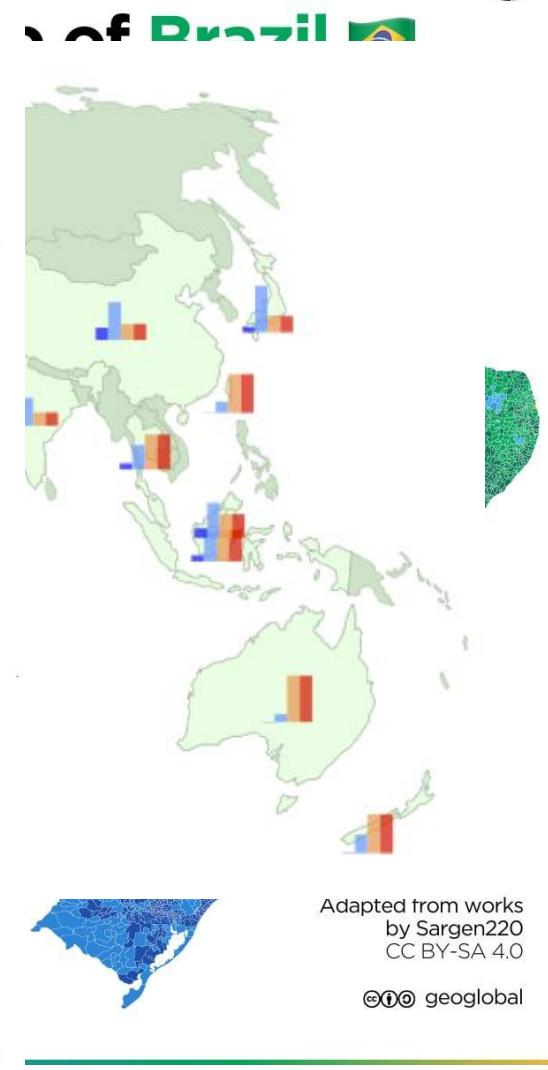
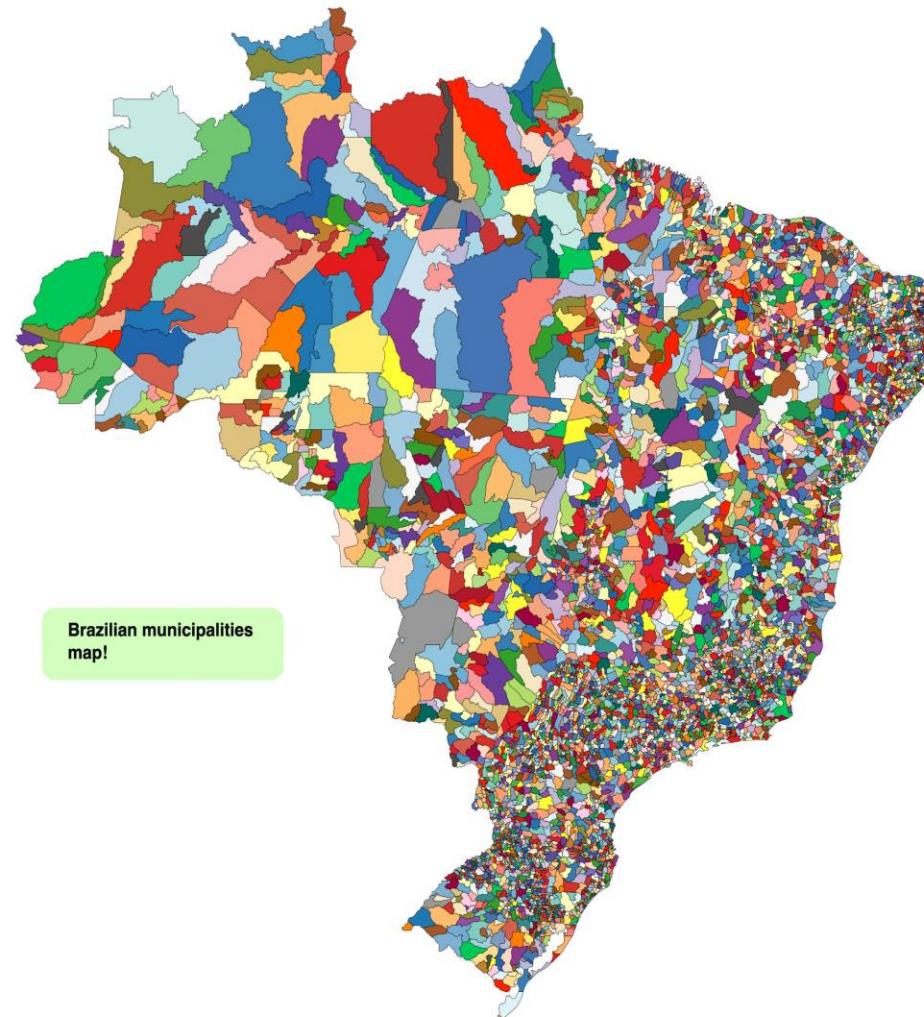
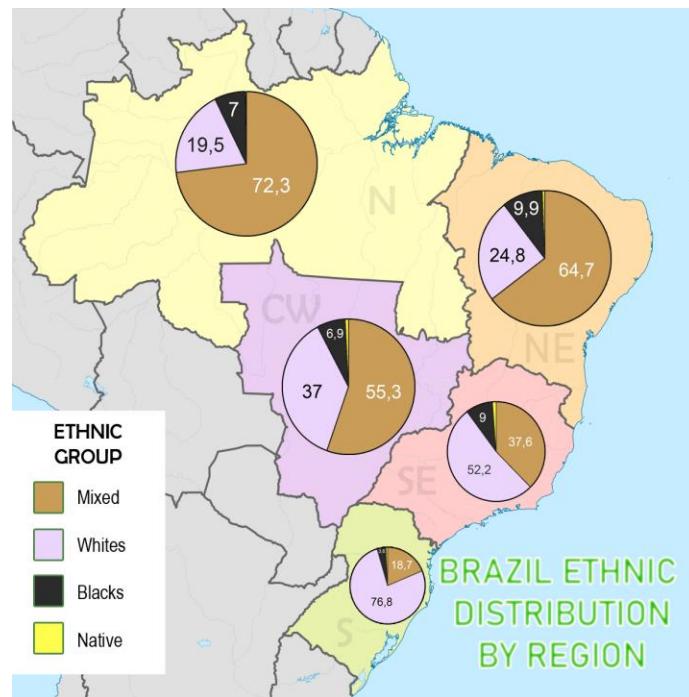


Visualização - gráficos de mapa

Valores numéricos por região

A se pensar:

- Cores ou marcadores?
- Tamanho dos marcadores
- Escala do mapa



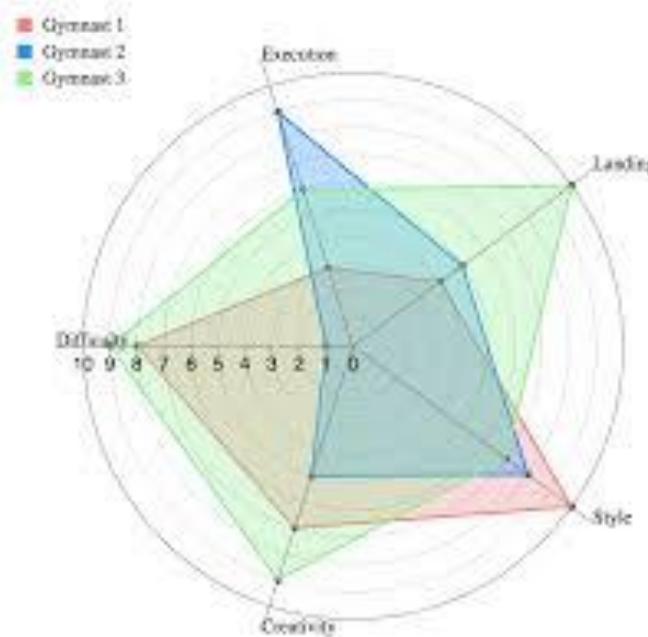
Visualização - gráficos de radar

Valores ordinais e numéricos por dois categóricos

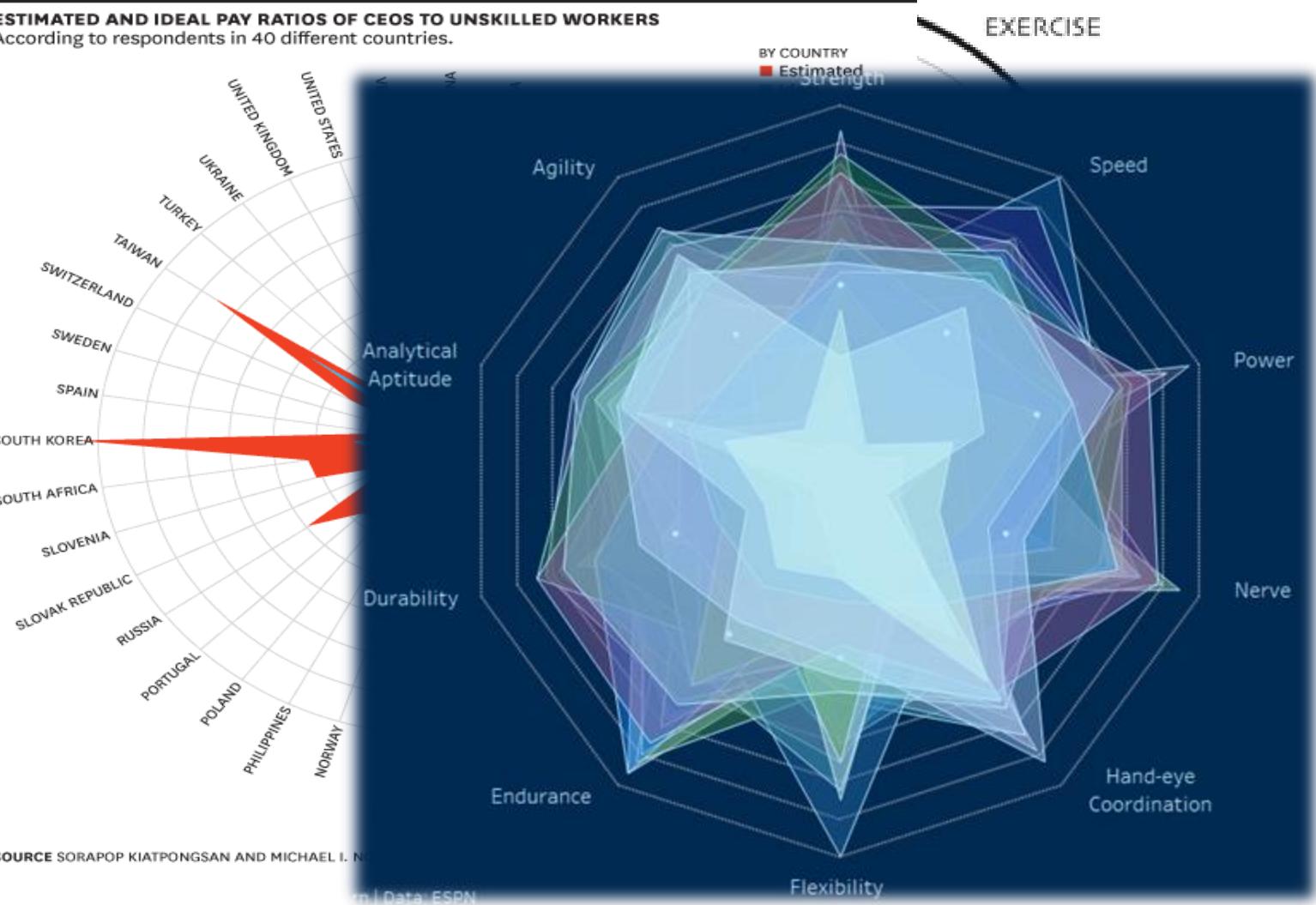
A se pensar:

- Espetos ou barras
- Quantidade de atributos
- Quantidade de classes

Gymnast Scoring Radar Chart



ESTIMATED AND IDEAL PAY RATIOS OF CEOS TO UNSKILLED WORKERS
According to respondents in 40 different countries.

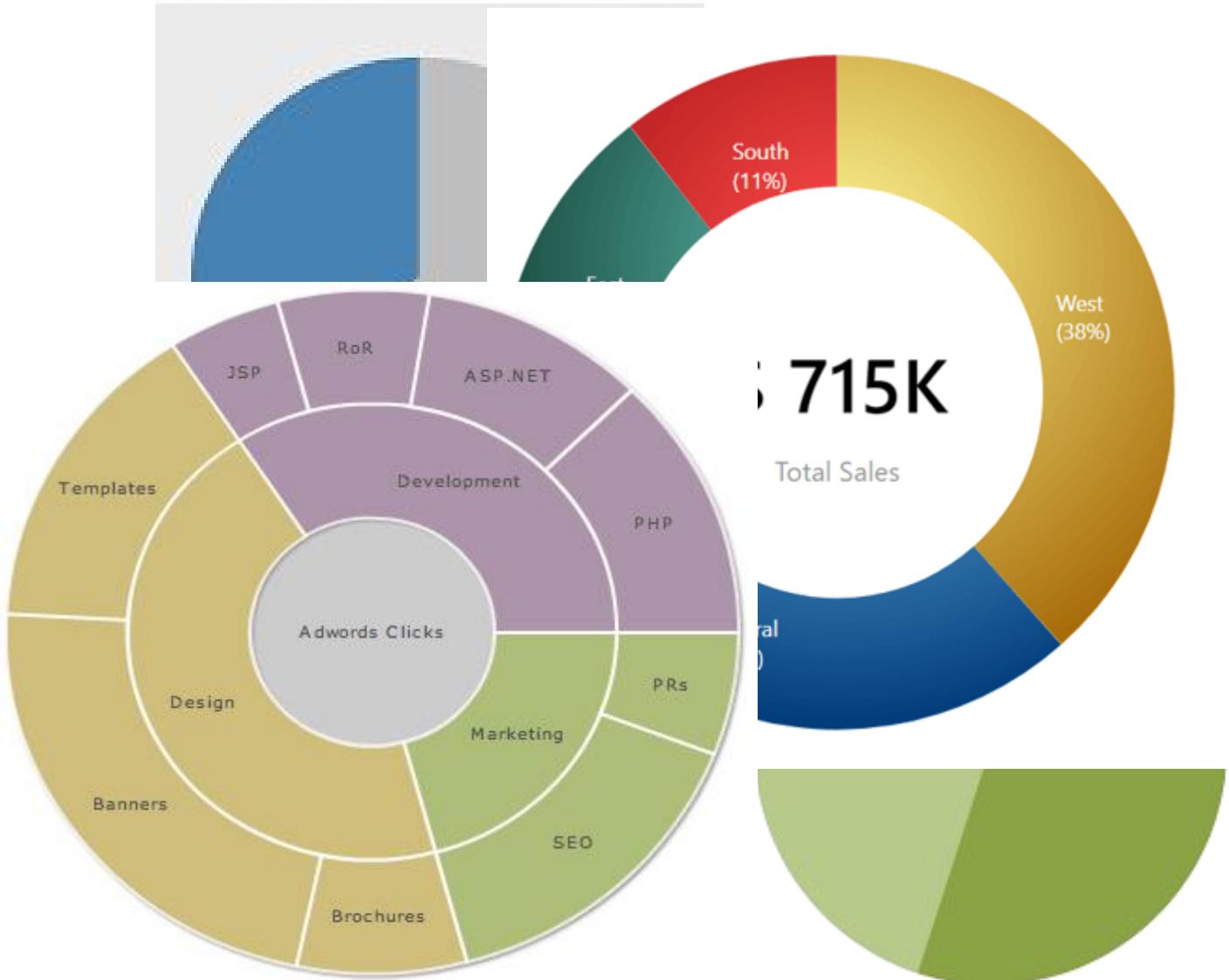
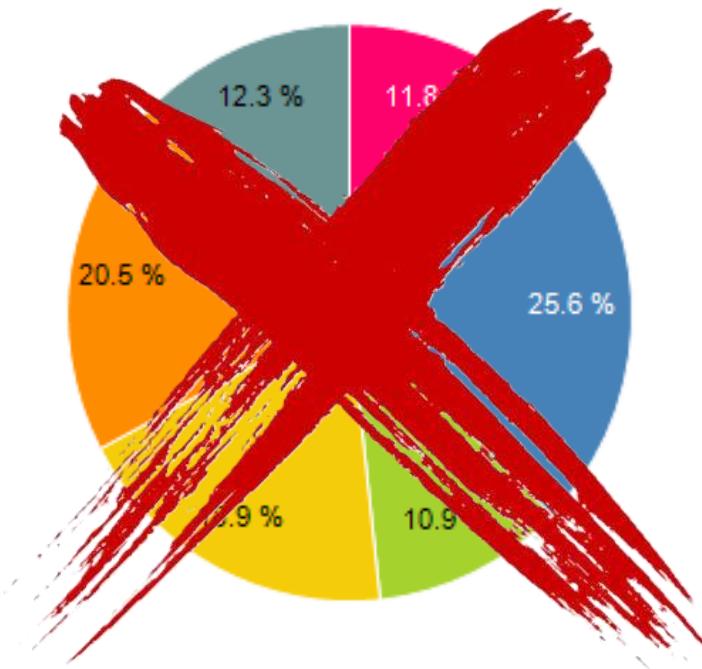


Visualização - gráficos de pizza

Valores numéricos por classes

A se pensar:

- Poucas ~~duas~~ classes
- Rosca ou pizza?
- Níveis de classe



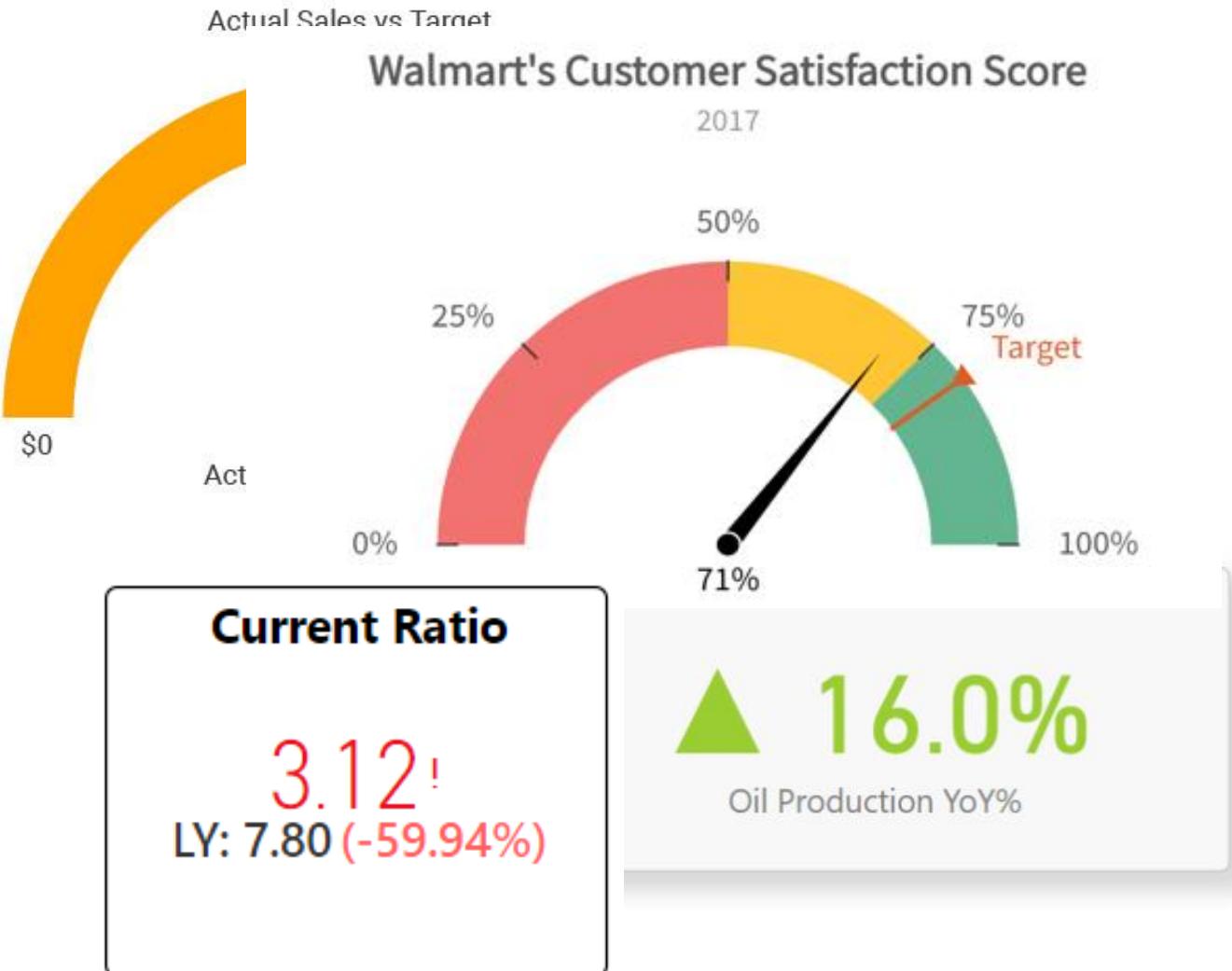
Valores numéricicos agregados

A se pensar:

- Cartão ou medidor?
- Tem referência, alvos, limites ou limiares de aceitação?

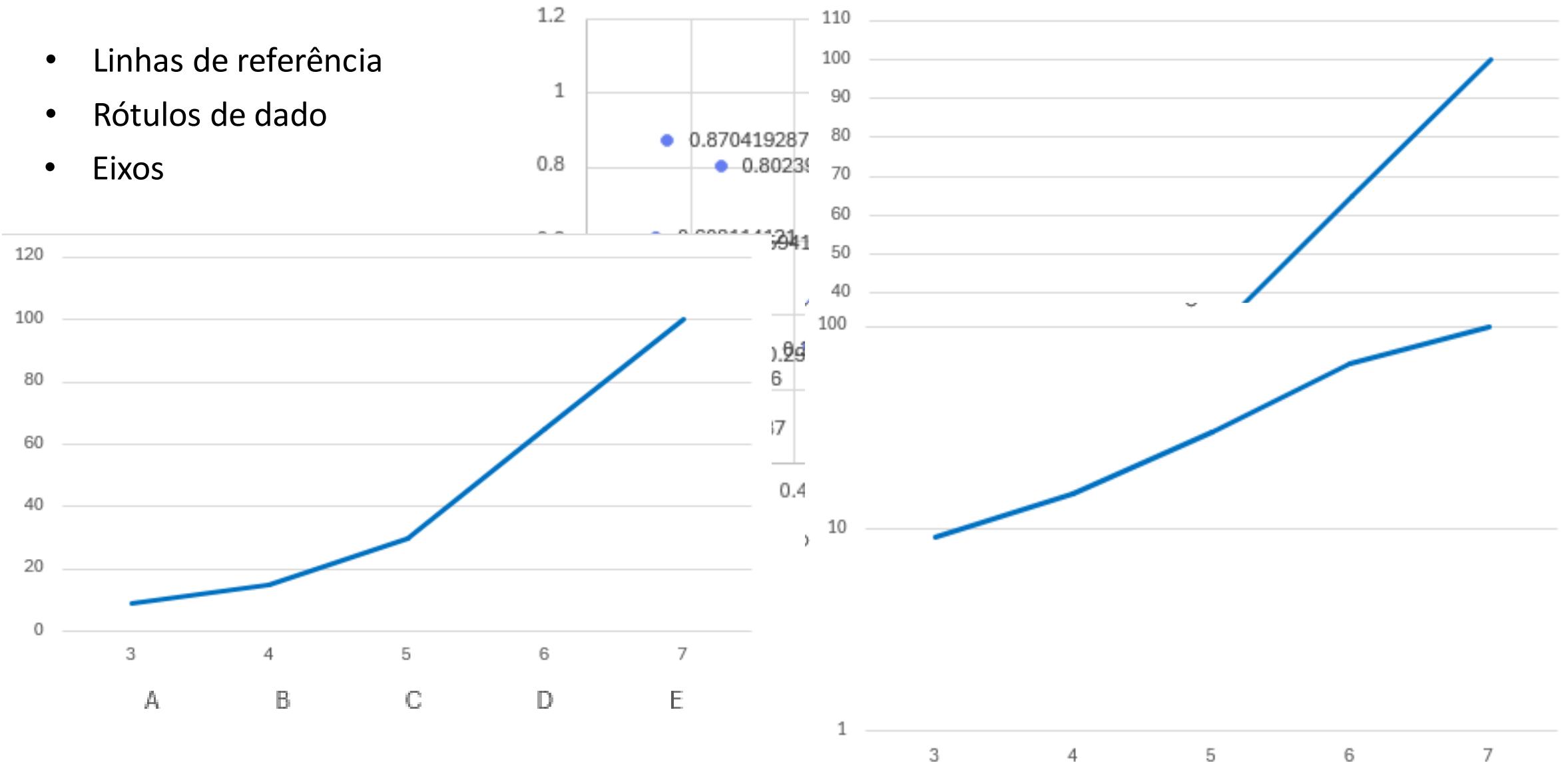
Lead Time Compra

18,2

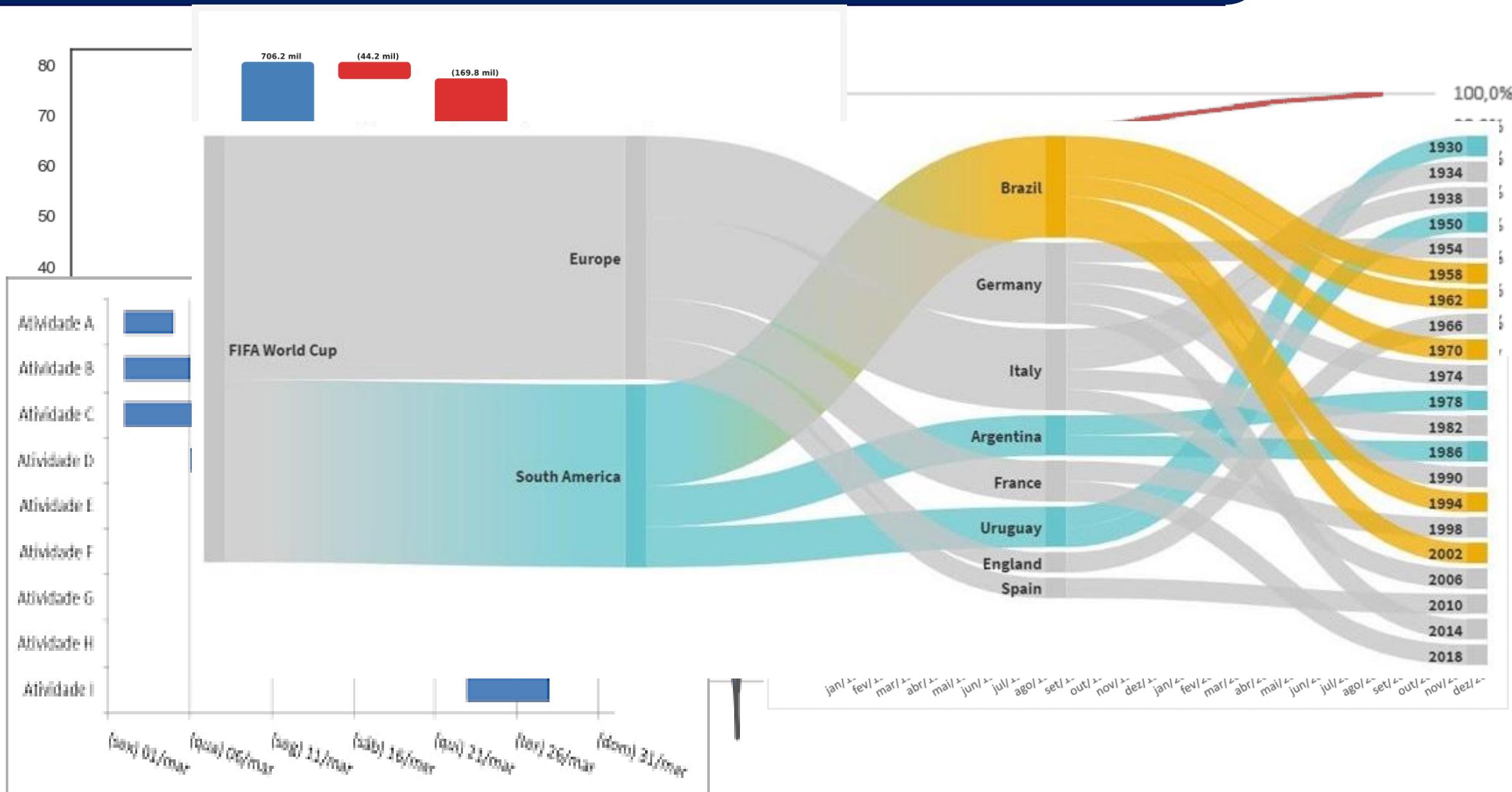


Visualização - cuidados gerais

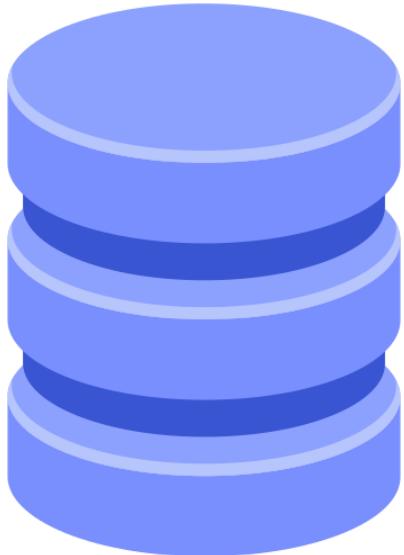
- Linhas de referência
- Rótulos de dado
- Eixos



Visualização - visualizações avançadas



Dados à obra!



Fontes

As tabelas resultado da camada anterior, com os dados criados com relevância e significado.

Cada uma das tabelas deve ser importada e corretamente relacionada para replicar as conexões feitas da camada anterior.

Saídas

Agora é a hora de botar a criatividade pra jogo.

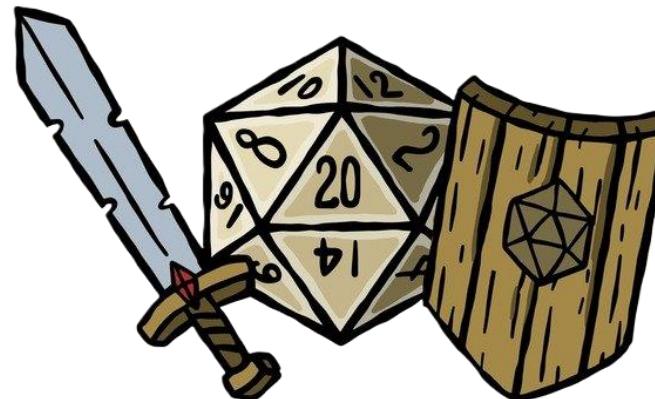
Conforme todos os princípios que estudamos, monte um relatório no PowerBI que mostre a relação entre os dados coletados na sua máquina e os dados meteorológicos da sua cidade.

Explore os diferentes tipos de dado e visuais para buscar praticidade e naturalidade com eles.

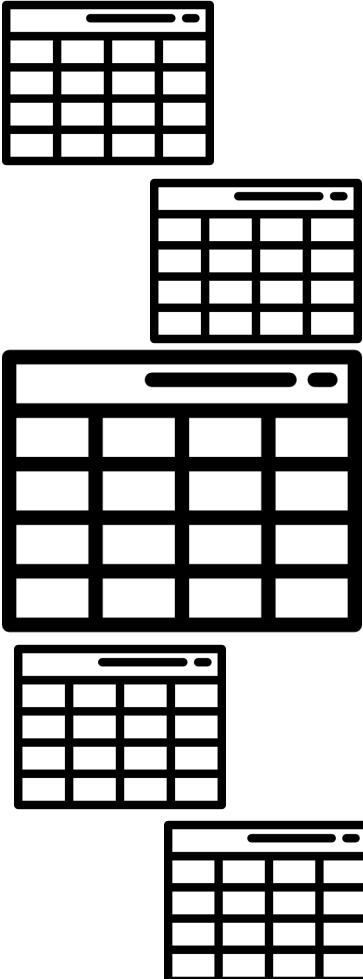


AULA 6

INTELIGÊNCIA



Camada ouro

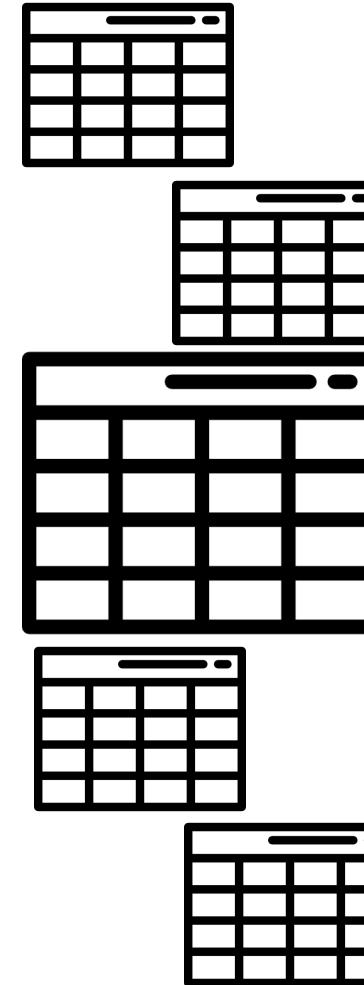


Camada elaborada com métodos matemáticos avançados e complexos.

Visa identificar padrões ocultos dentro dos dados disponíveis de forma a agregar e facilitar os processos dos usuários.

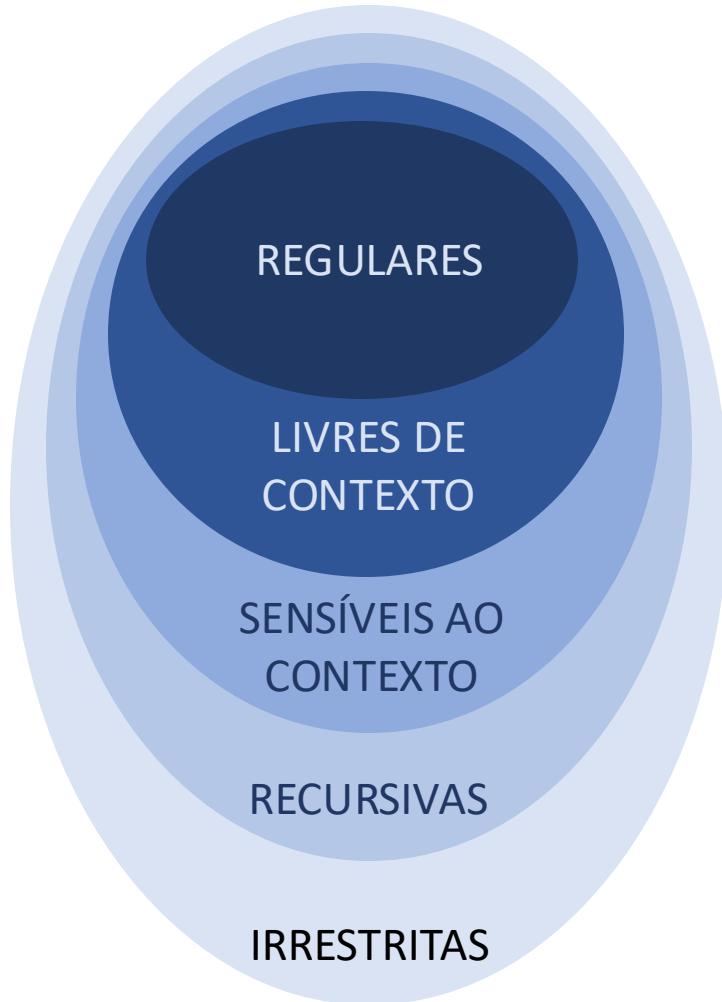
É a etapa mais empírica e imprevisível do processo.

Camada ouro



O que é Inteligência Artificial?

Hierarquia/taxonomia de Chomsky para linguagens



O estudo das linguagens se refere ao estudo da viabilidade da prova da veracidade de sentenças, ou seja, resolução de problemas matemáticos. Chomsky classificou as linguagens em 4 tipos.

Regulares: seguem padrões fixos de repetição sequencial (REGEX)

Livres de contexto: seguem regras de derivação simples, onde a validade de uma palavra depende apenas do contexto diretamente anterior à ele (outras linguagens de programação no geral)

Sensíveis ao contexto: tem regras de derivação mais complexas, onde a validade de uma palavra depende de um contexto amplo (a maior parte dos problemas resolvidos por computação tradicional)

Recursivas: conjunto de todas as linguagens aceitáveis por uma Máquina de Turing (qualquer computador tradicional)

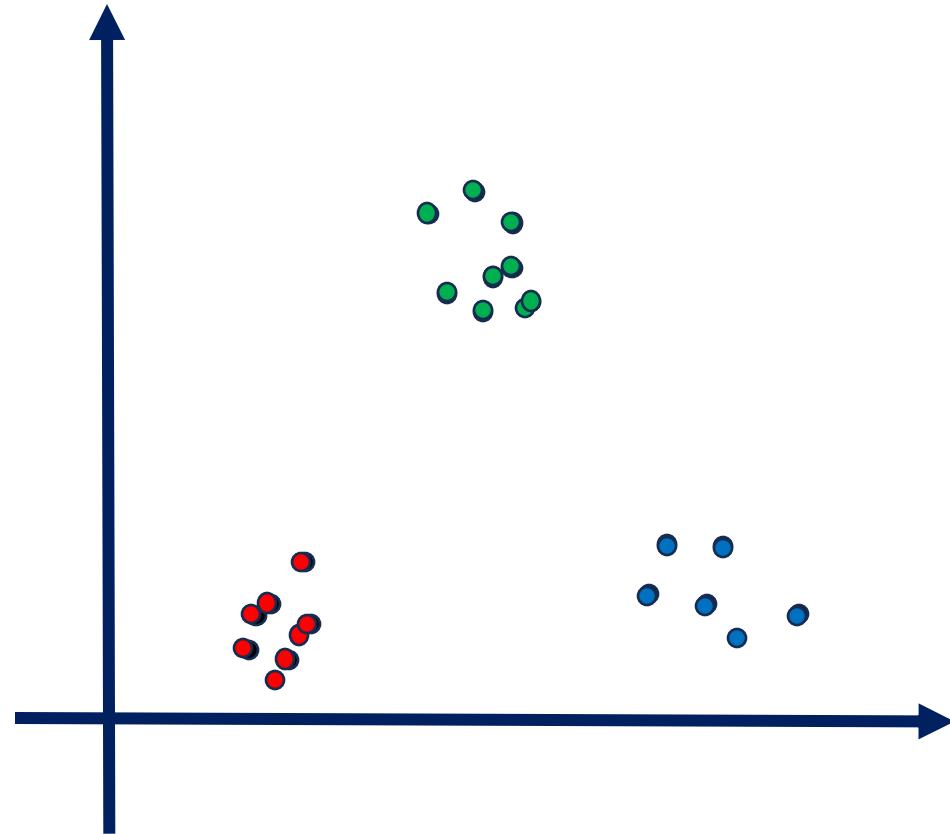
Irrestritas: qualquer linguagem sem restrições de formatos de derivação (línguas faladas)



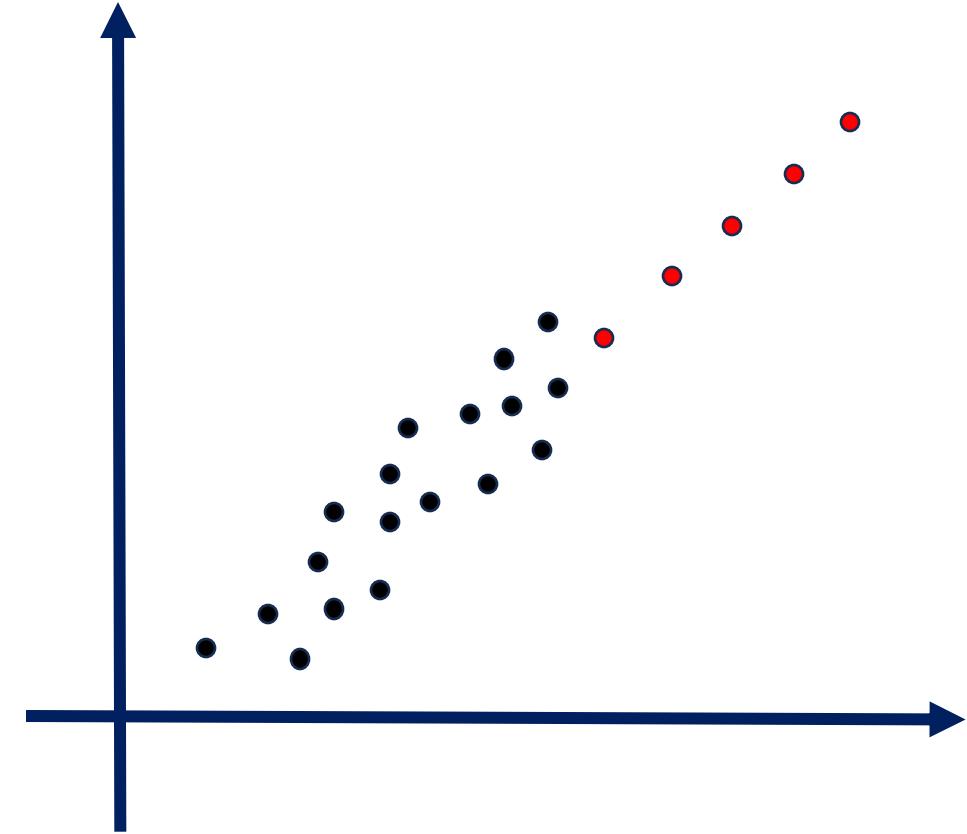
Inteligência Artificial

Tipos de problemas

Classificação



Régressão



Ciclo de criação



Pro

Etapa de construção do modelo e refinamento dos seus parâmetros para se adequar aos dados.

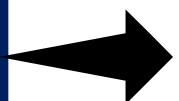
Disponibilização do modelo para uso e consumo pelos usuários. É necessária atenção periódica para solução de problemas de performance.

Pré-
processamento

Definição de
linha de base

Treino

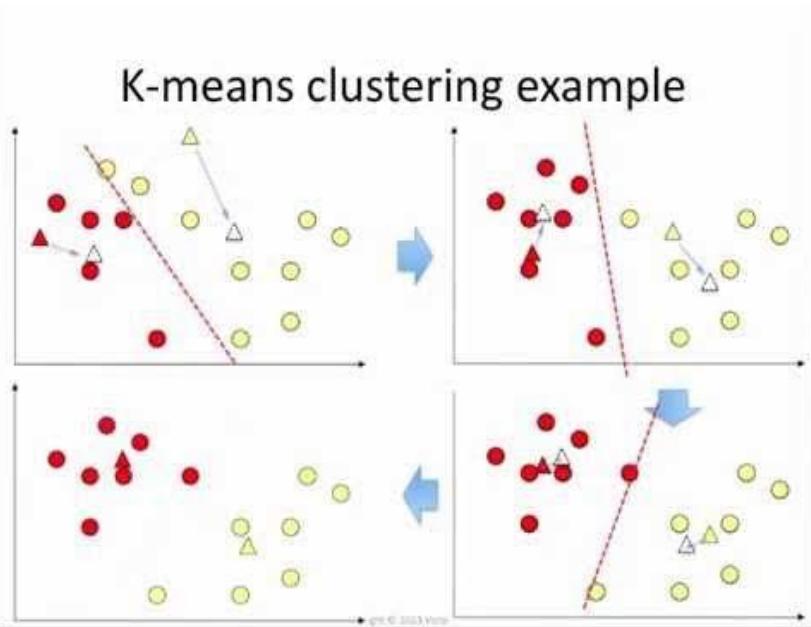
Deploy e
manutenção



Métodos de classificação

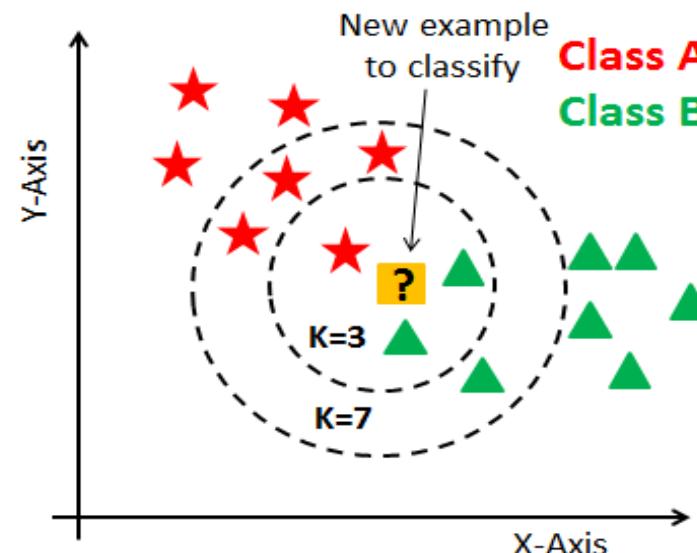
K-means

K-means clustering example



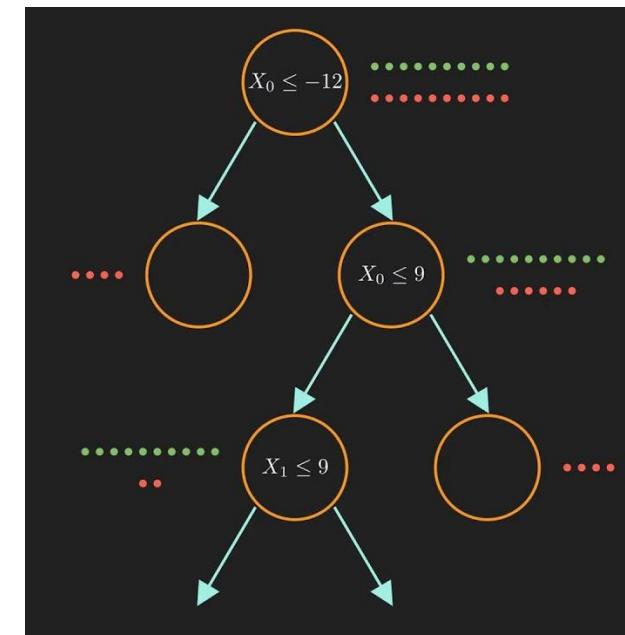
Processo iterativo onde os centros de cada um dos k grupo são reajustados de acordo com as amostras dele em cada iteração.

KNN



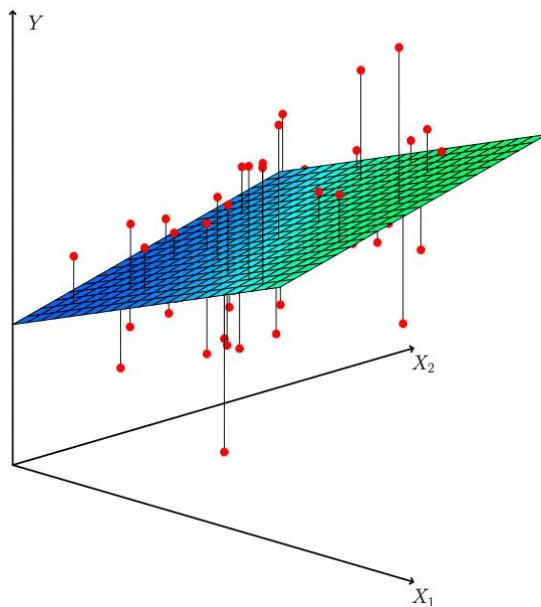
Novas amostras são classificadas com base nos k vizinhos mais próximos dela.

Árvore/floresta de decisão



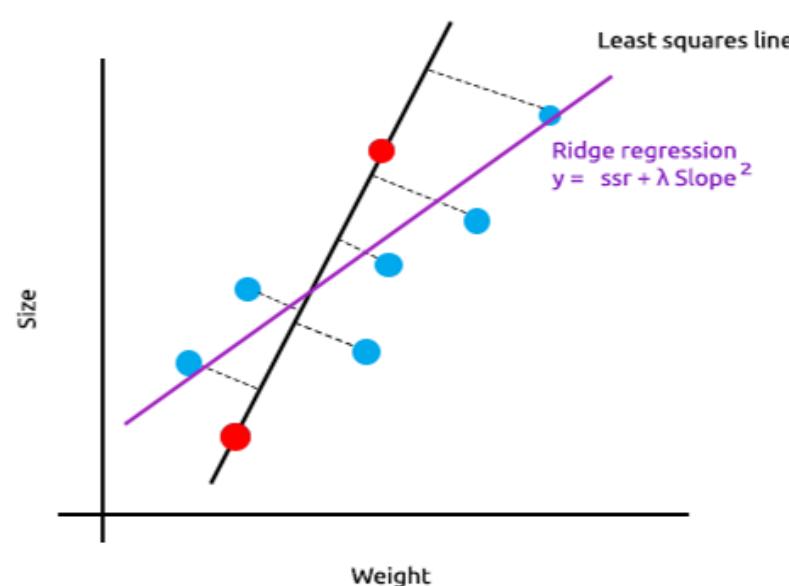
As amostras são recursivamente divididas de acordo com suas propriedades criando um classificador determinístico.

Regressão Linear



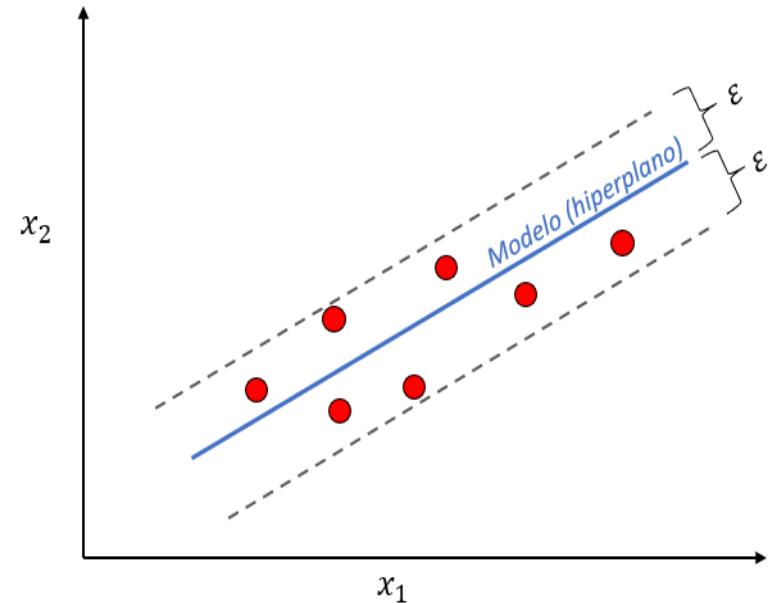
Estima-se uma reta que prevê o comportamento da saída dadas as entrada, de acordo com a correlação entre seus valores.

Regressão Ridge/Lasso



Similar à regressão linear, mas limita os valores dos coeficientes para que as entradas não tenham pesos muito discrepantes.

Regressão de Vetor de Suporte



É criado um hiperplano passando através dos dados de forma que o erro máximo entre o hiperplano e os dados seja limitado à uma constante.

Métricas de avaliação - classificação

Matriz de confusão

	Cão	Gato	Coelho
Cão	88	14	4
Gato	12	85	11
Coelho	5	15	91
	Cão	Gato	Coelho

Acuidade

O total de acertos do

total de acertos entre o que foi
considerado pertencente à classe

total de acertos entre o que é
realmente pertencente à classe

f1

mediação entre a precisão e a
revocação

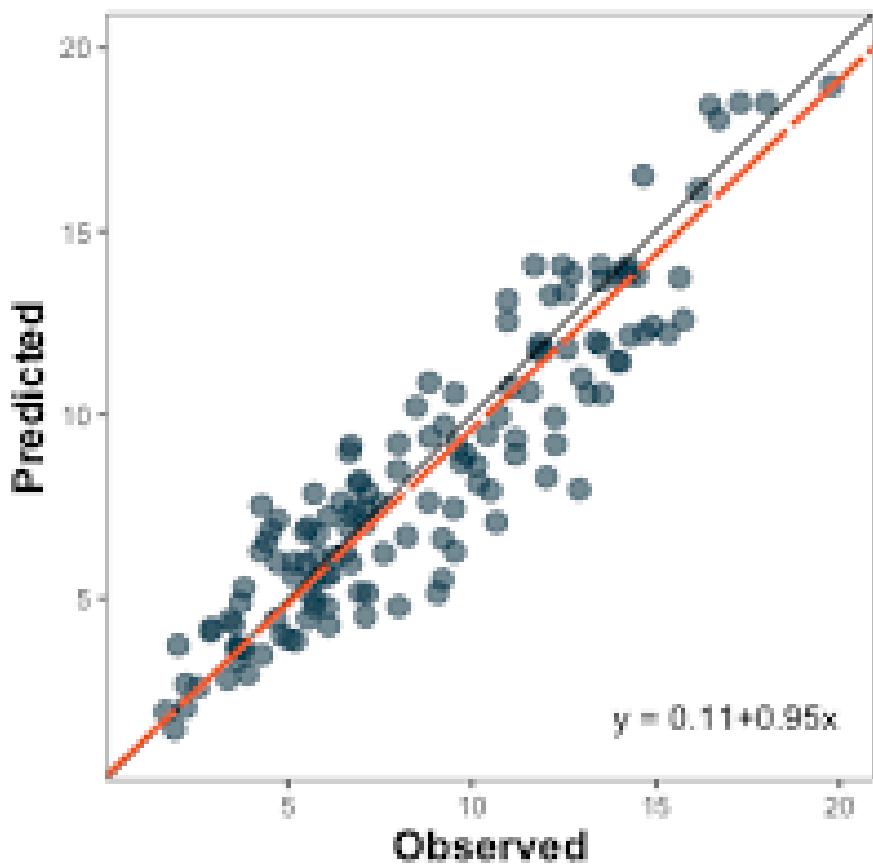
$$\frac{VP + VN}{VP + VN + FP + FN}$$

$$\frac{VP}{VP + FP}$$

$$\frac{VP}{VP + FN}$$

$$\frac{1}{\frac{1}{Prec} + \frac{1}{Rev}}$$

Gráfico real X predito



Erro Médio Absoluto

A média do total de divergência entre os valores preditos e esperados

Erro Quadrático Médio

A média do total de divergência corrigido para pequenos valores

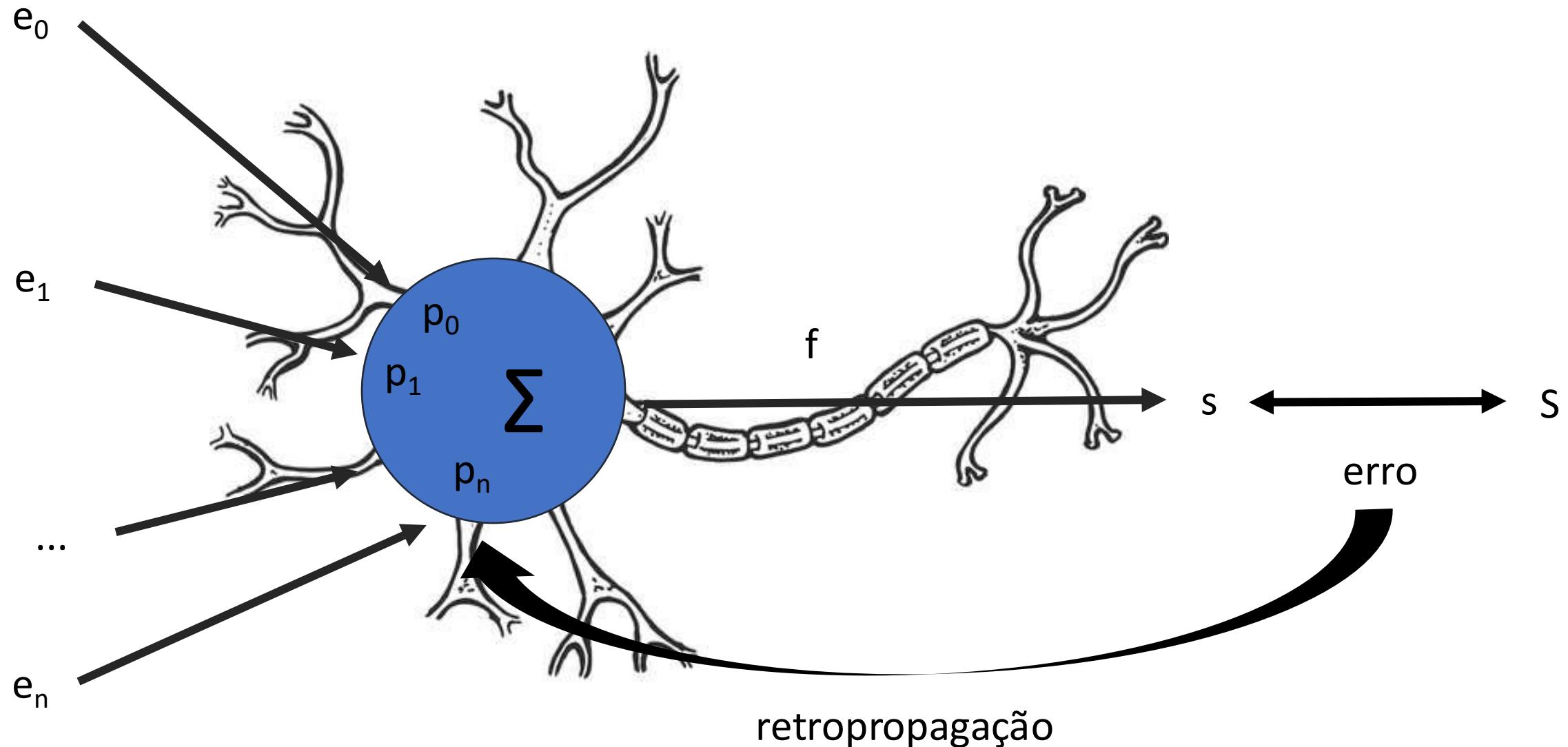
R^2

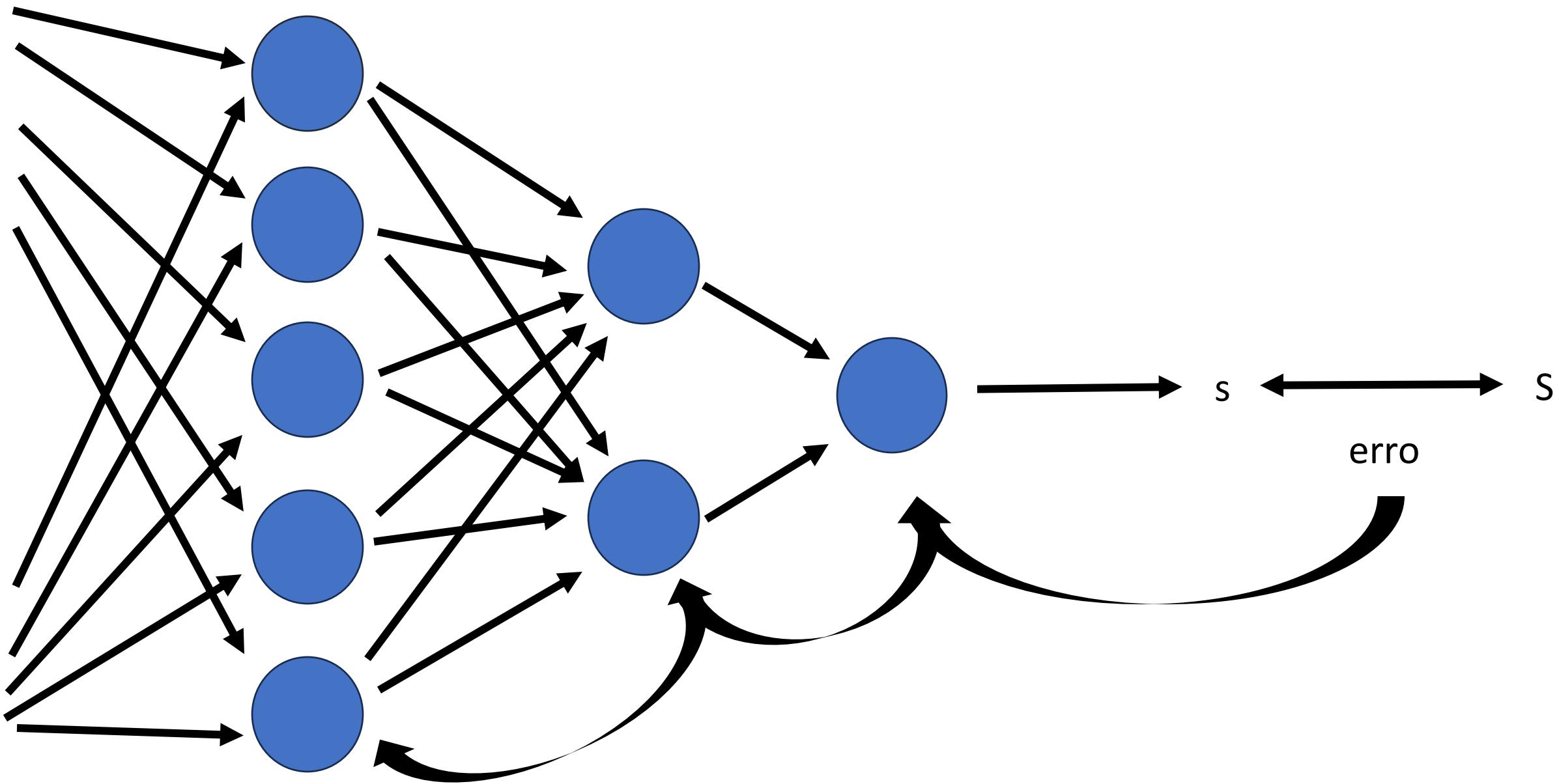
Representa quanto do erro do modelo está contido na variância original da amostra

Variância explicada

Representa o quanto da variância do erro é proveniente da variância da amostra

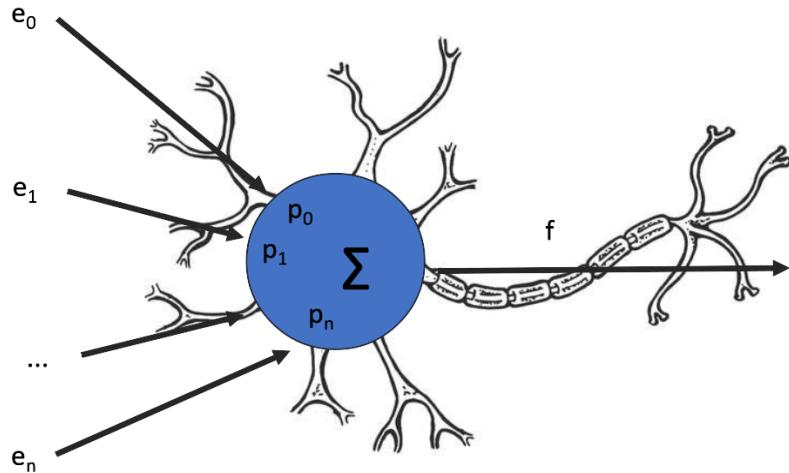
Redes Neurais - perceptron



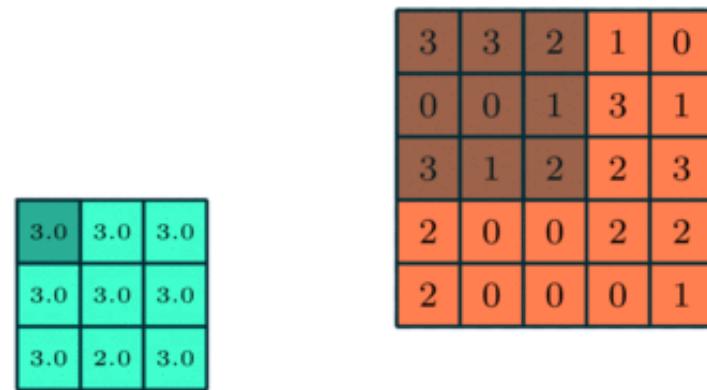


Tipos de neurônio

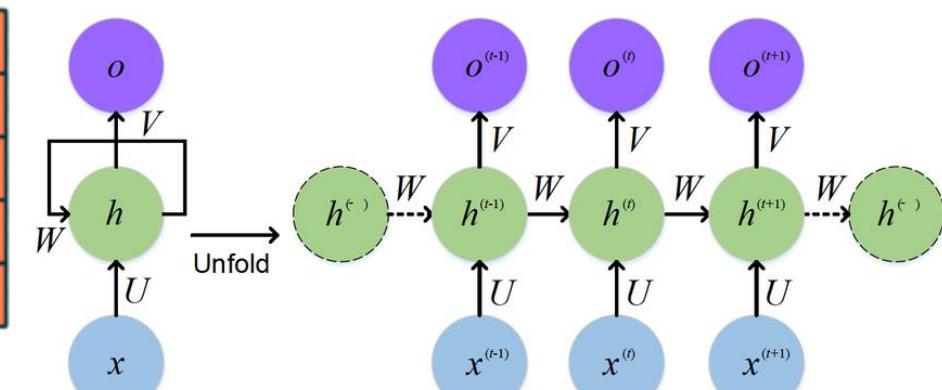
Perceptron



Convolucional



Recorrente



Tipo básico de neurônio que aplica soma ponderada. É muito útil para tipos tabulares de dados com muitas colunas.

Aplica filtro matricial sobre as entradas. Focada em processamento de imagens, pois esses filtros identificam padrões nos dados de entrada.

Utiliza a própria saída como uma nova entrada entre as iterações de treino. Facilita tratamento de dados temporais.

Muitos anos depois, diante do pelotão de fuzilamento, o Coronel Aureliano Buendía havia de recordar aquela tarde remota em que seu pai o levou para conhecer o gelo. Macondo era então uma aldeia de vinte casas de barro e taquara, construídas à margem de um rio de águas diáfanas que se precipitavam por um leito de pedras polidas, brancas e enormes como ovos pré-históricos. O mundo era tão recente que muitas coisas careciam de nome e para mencioná-las se precisava apontar com o dedo. Todos os anos, pelo mês de março, uma família de ciganos esfarrapados plantava a sua tenda perto da aldeia e, com um grande alvoroço de apitos e tambores, dava a conhecer os novos inventos. Primeiro trouxeram o imã. Um cigano corpulento, de barba rude e mãos de pardal, que se apresentou com o nome de Melquíades, fez uma truculenta demonstração pública daquilo que ele mesmo chamava de a oitava maravilha dos sábios alquimistas da Macedônia.

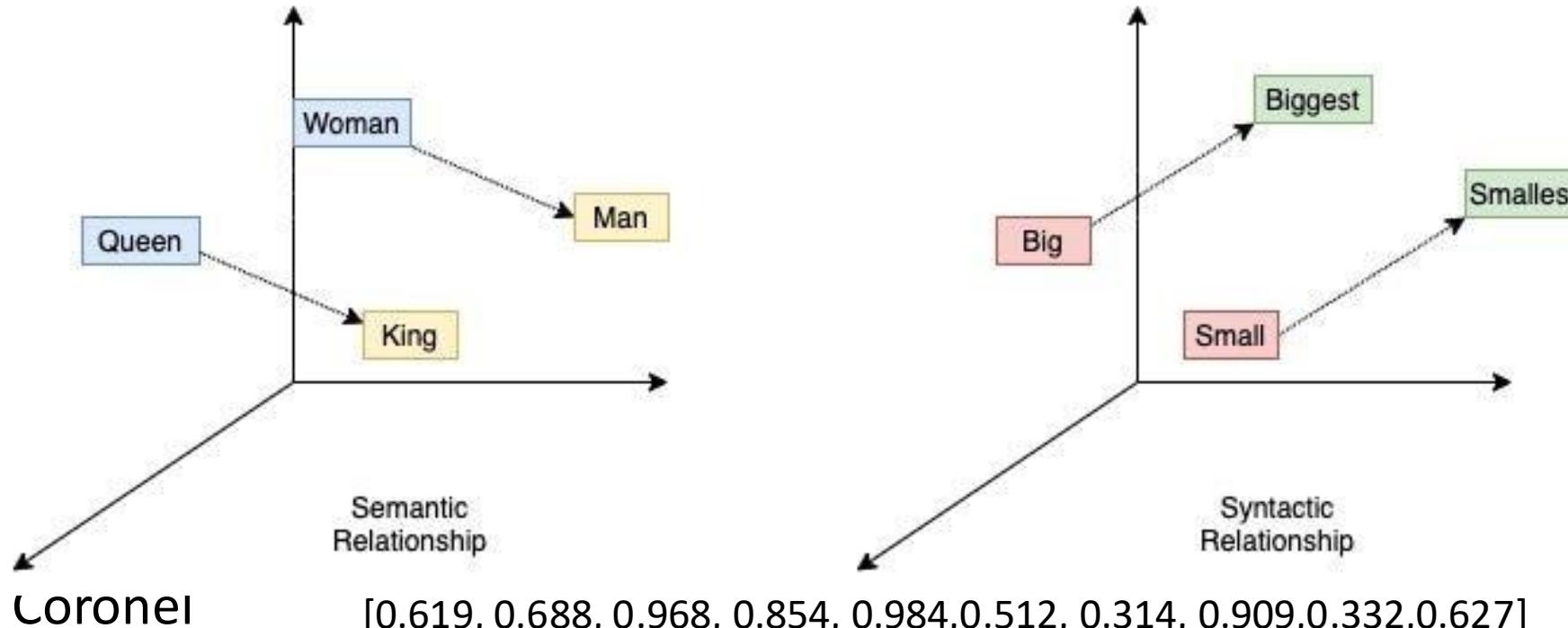
Tokenização

Muitos anos depois, diante do pelotão de fuzilamento, o Coronel Aureliano Buendía havia de recordar aquela tarde remota em que seu pai o levou para conhecer o gelo. Macondo era então uma aldeia de vinte casas de barro e taquara, construídas à margem de um rio de águas diáfanas que se precipitavam por um leito de pedras polidas, brancas e enormes como ovos pré-históricos. O mundo era tão recente que muitas coisas careciam de nome e para mencioná-las se precisava apontar com o dedo. Todos os anos, pelo mês de março, uma família de ciganos esfarrapados plantava a sua tenda perto da aldeia e, com um grande alvoroço de apitos e tambores, dava a conhecer os novos inventos. Primeiro trouxeram o imã. Um cigano corpulento, de barba rude e mãos de pardal, que se apresentou com o nome de Melquíades, fez uma truculenta demonstração pública daquilo que ele mesmo chamava de a oitava maravilha dos sábios alquimistas da Macedônia.

Tokonização

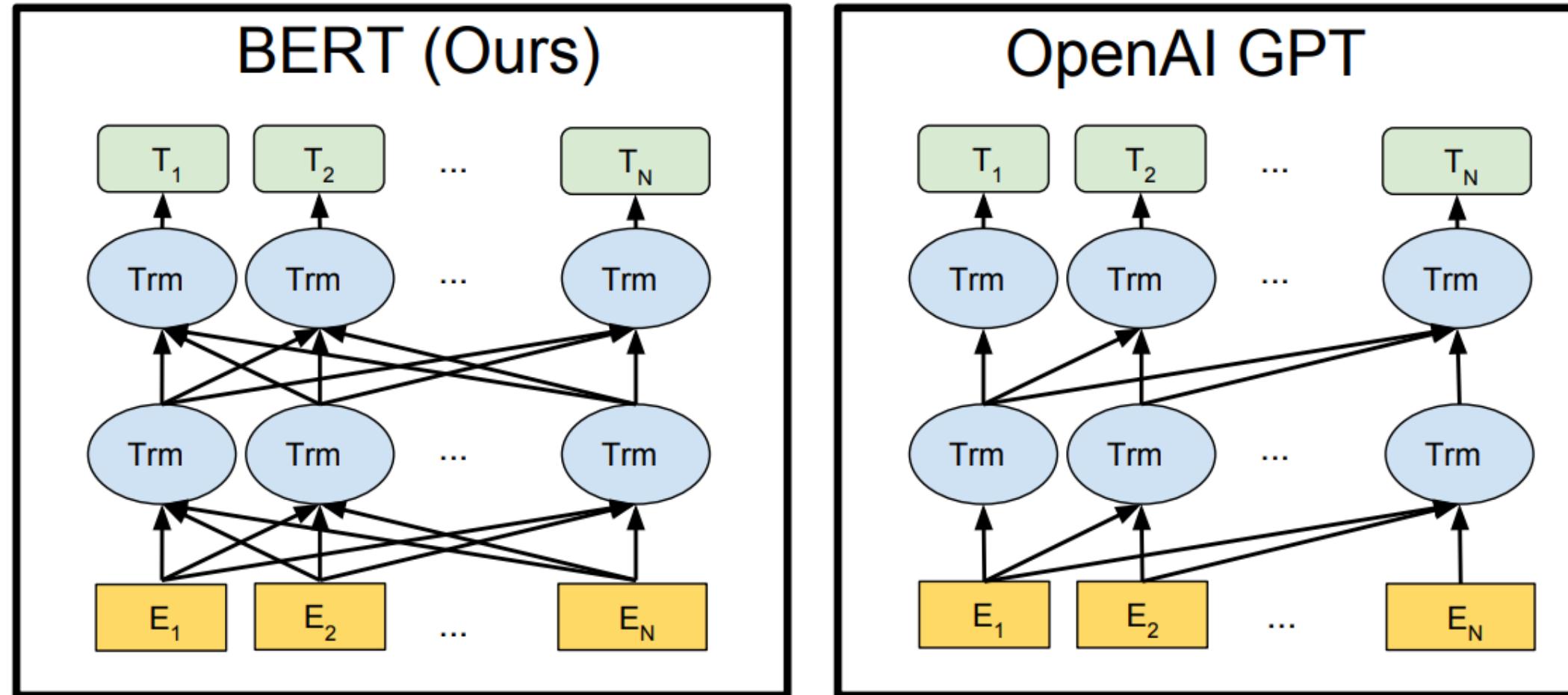
anos

[0.008, 0.518, 0.295, 0.246, 0.063, 0.04 , 0.67 , 0.227, 0.505,0.676]



recordar

[0.761, 0.915, 0.207, 0.466, 0.004, 0.985, 0.417, 0.398, 0.454, 0.625]

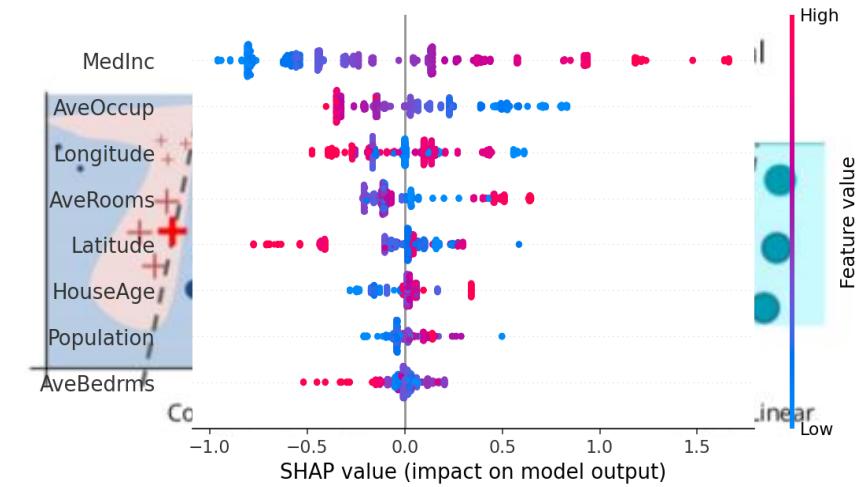
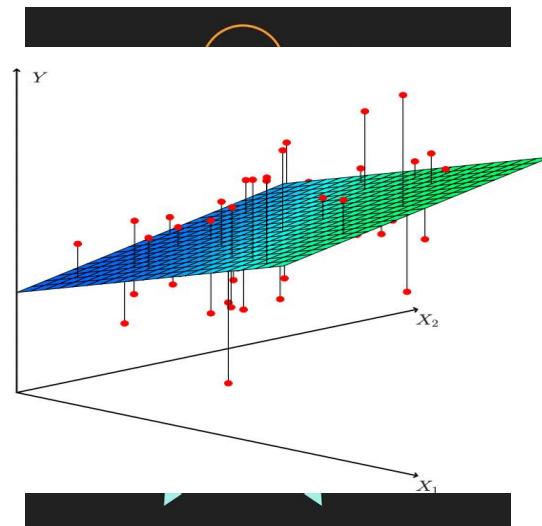


amento

[0.761, 0.915, 0.207, 0.466, 0.007, 0.04, 0.985, 0.5417, 0.398, 0.454, 0.625]

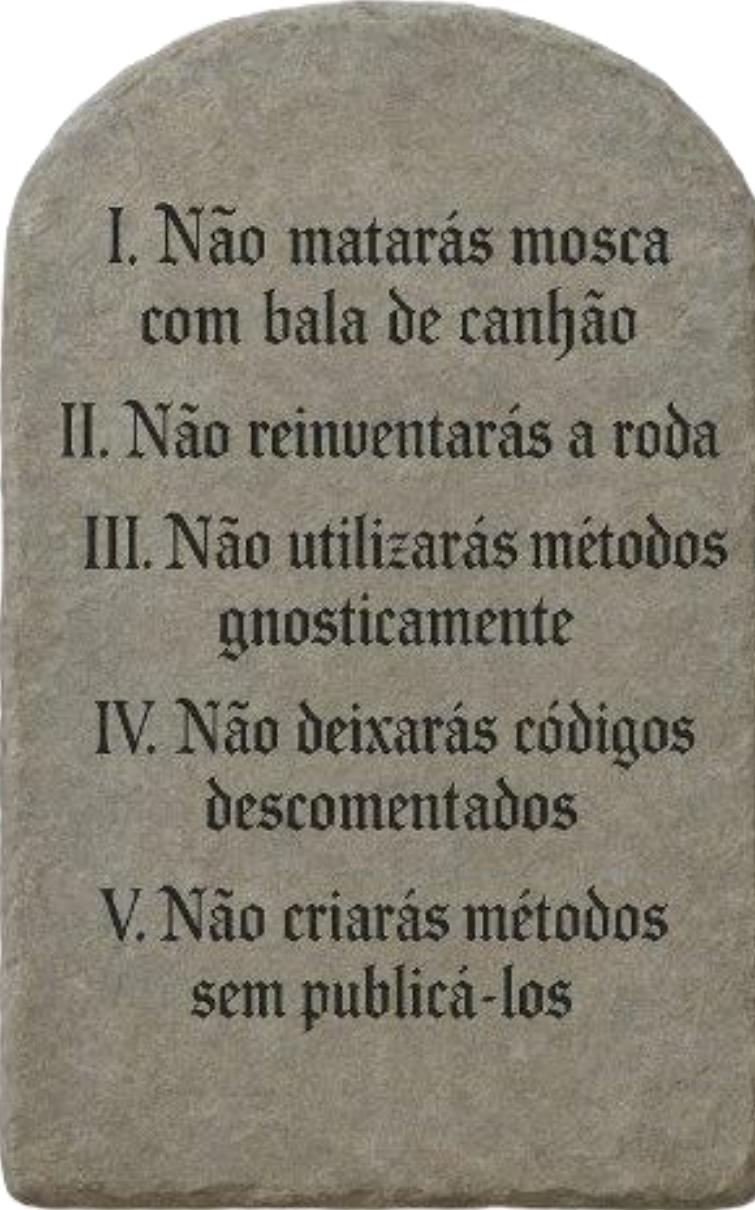
Interpretabilidade

As alegações de interpretação são sempre feitas, os interpretadores são profissionais que integram e gerenciam as demandas de saída das classes apresentadas pelo modelo.



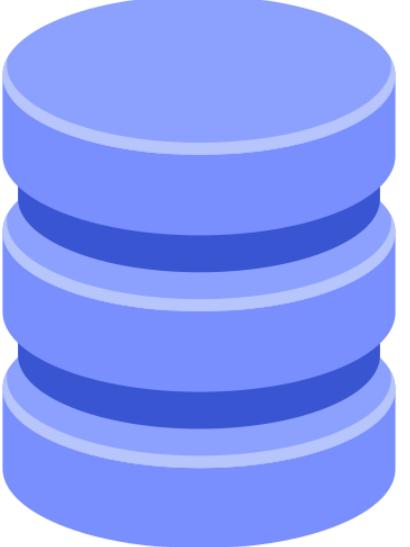
Os modelos lineares são explicáveis, os resultados de modelagem são transparentes e passam quanto quais variáveis influem na obtenção dos resultados obtidos. Os resultados obtidos em regressões lineares.

Explicabilidade

- 
- I. Não matarás mosca com bala de canhão
 - II. Não reinventarás a roda
 - III. Não utilizarás métodos gnosticamente
 - IV. Não deixarás códigos descomentados
 - V. Não criarás métodos sem publicá-los

Dados à obra!

Fontes



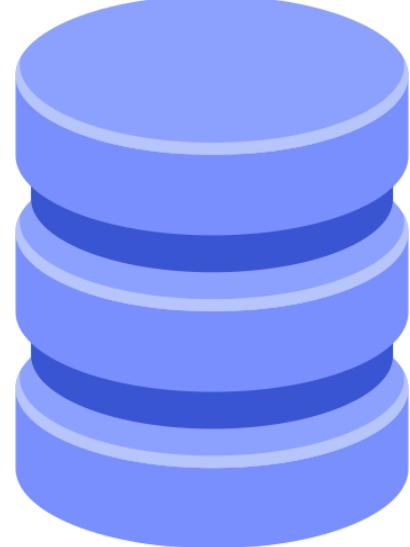
As tabelas resultado da camada ouro, com os dados criados com relevância e significado.

Cada uma das tabelas deve ser importada e corretamente relacionada para replicar as conexões feitas da camada anterior.

Saídas

Agora vamos encontrar novas relações entre os dados.

Tente implementar e treinar algoritmos de Inteligência Artificial sobre os dados para entender quais fatores são mais críticos para o aumento de temperatura da CPU do seu computador ou qualquer outro valor que você tenha medido dela.



Não se limite aos modelos apresentados aqui, busque e proponha novas possibilidades.

AVALIAÇÃO





Obrigado!

Acesse o site pelo QR Code



contato@scalait.com



www.scalait.com



www.linkedin.com/scalait



@scala.it