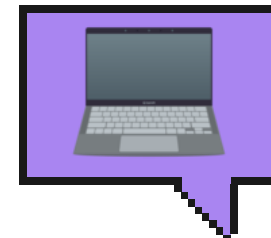


Introdução à Preparação e Análise de Dados



PUCRS

Pontifícia Universidade Católica
do Rio Grande do Sul



- Aula 11 -
Coleta, Preparação e
Análise de Dados

Prof. Me. Lucas R. C. Pessutto

Slides adaptados do material do Prof. Lucas Silveira
Kupssinskü e do Prof. Luan Fonseca Garcia

Análise de Dados

- Para conduzir uma mineração ou análise de dados de sucesso é **essencial** que conheçamos nossos dados.
 - Quais **tipos** de **atributos** ou **campos**?
 - Que **tipos** de **valores** cada atributo tem?
 - Como esses valores estão **distribuídos**?
 - Como podemos medir a **similaridade** entre determinados dados em relação a outros?
- Este tipo de insight facilita a análise subsequente dos dados!

Um exemplo...

012	232	33.5	0	10.7
020	121	16.9	2	210.1
027	165	24.0	0	427.6
⋮				

Data Objects

- Data sets são compostos de objetos de dados (data object).
- Um **objeto de dado** representa uma entidade.
 - Mesmo que amostra, exemplo, instância, ponto, tupla, objeto.
- Exemplos:
 - Banco de vendas: clientes, vendas, itens da loja;
 - Base médica: pacientes, tratamentos;
 - Banco da universidade: estudantes, professores, cursos
- Dados são descritos por **atributos**.
- Linhas em um BD-> objetos de dado; Colunas -> atributos

Dados costumam vir em diferentes sabores

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	class
0	25	Private	226802	11th	7	Never-married	Machine-op-inspct	Own-child	Black	Male	0	0	40	United-States	<=50K
1	38	Private	89814	HS-grad	9	Married-civ-spouse	Farming-fishing	Husband	White	Male	0	0	50	United-States	<=50K
2	28	Local-gov	336951	Assoc-acdm	12	Married-civ-spouse	Protective-serv	Husband	White	Male	0	0	40	United-States	>50K
3	44	Private	160323	Some-college	10	Married-civ-spouse	Machine-op-inspct	Husband	Black	Male	7688	0	40	United-States	>50K
4	18	?	103497	Some-college	10	Never-married	?	Own-child	White	Female	0	0	30	United-States	<=50K

Atributos

- **Atributo (ou dimensão, features, variável):** um campo do banco de dados, representando alguma característica do objeto de dado.
 - *Ex.: cliente_ID, nome, endereço*
- Tipos:
 - Nominal
 - Binário
 - Ordinal
 - Numérico (quantitativo)
 - Em escala intervalar
 - Em escala proporcional

Tipos de Atributos Qualitativos

- **Nominal:** categorias, estados, “nomes de coisas”
 - *Cor_do_cabelo* = {loiro, preto, branco, castanho}
 - Estado civil, ocupação, identidade, CEP
- **Binário**
 - Atributo nominal com apenas dois estados (0 e 1)
 - Simétricos: valores com importância igual.
 - Ex: macho ou fêmea
 - Assimétricos: valores com importância distinta.
 - Teste médico (positivo vs. negativo)
 - Convenção: atribuir 1 para o valor mais importante (tem HIV)
- **Ordinal**
 - Valores possuem uma ordenação significativa (ranqueamento), mas a magnitude entre os valores não é conhecida.
 - *Tamanho* = {pequeno, médio, grande}

Tipos de Atributos Numéricos

- Quantidades (inteiros ou valores reais)
- **Intervalar**
 - Medidos em uma escala em que **unidades possuem o mesmo tamanho**
 - Valores possuem ordenamento
 - *Temperatura em C° ou F°, Datas no calendário*
 - Não há um “ponto zero” real, não podemos estabelecer proporções
- **Proporcional**
 - Possuem um **ponto zero** inerente real
 - Podemos calcular proporções entre os valores
 - 10 K° é o dobro do que 5 K°.
 - 2 metros é o dobro de 1 metro.
 - Comprimento, dinheiro, “contagens” em geral (anos de experiência, número de palavras, etc)

Representação numérica de atributos

- Podemos representar atributos nominais utilizando **números**.
 - Classe 1, classe 2, classe 3.
 - 0 ou 1 para binários.
- Isto não faz deles atributos numéricos!
- Não há uma noção de escala entre os valores, é apenas a representação!

Atributos Discretos vs. Contínuos

- **Atributo Discreto**

- Possui um conjunto de valores possíveis “contável” (finito)
 - CEP, profissões, conjunto de palavras em um doc.
- “**Contável infinito**” é quando os possíveis valores são infinitos, mas na prática temos uma relação de um para um com números naturais. Ex: id_cliente

- **Atributo contínuo**

- Possui o conjunto dos reais como possíveis valores
 - Temperatura, altura, peso
- Obviamente na prática valores reais só podem ser medidos e representados utilizando um conjunto finito de dígitos.
- Tipicamente representados como variáveis de ponto flutuante.

Tipos de Dados

- Conforme tipo:
 - Numérico
 - Categórico
- Conforme operações:
 - Posso testar igualdade?
 - Existe uma relação de ordem?
 - Faz sentido adicionar valores?
 - Faz sentido multiplicar?

Resumo

Tipo		Descrição	Exemplo	Operações
Categórico (Qualitativo)	Nominal	Um atributo nominal só nos possibilita saber se um objeto é igual ou diferente a outro.	Cep, cpf, ids, cor dos olhos.	Igualdade
	Binário	Apenas dois valores possíveis	Masculino/Feminino	Igualdade
	Ordinal	Possui uma noção de ordem.	Tamanho, Escala Likert.	Comparação
Numérico (Quantitativo)	Intervalar	Diferenças entre valores tem significado.	Temperatura em Celsius, datas.	Adição e Subtração
	Razão ou Proporção	Razões possuem significado.	Temperatura em Kelvin, quantidades monetárias, massa, idade.	Multiplicação e Divisão

Tipos de Transformações

Tipo		Transformação	Exemplo
Categórico (Qualitativo)	Nominal	Qualquer mapeamento um-para-um.	Permutação ou reindexação.
	Ordinal	Qualquer função que preserva ordem. $x^{(1)'} = f(x^{(1)})$ tal que f é uma função monotônica	A noção de ruim, regular e bom pode ser igualmente representada pelos valores $\{-1,0,1\}$.
Numérico (Quantitativo)	Intervalar	$x^{(1)'} = ax^{(1)} + b$, tal que a e b são constantes	Conversão de fahrenheit para celsius.
	Razão ou Proporção	$x^{(1)'} = ax^{(1)}$	Um valor monetário pode ser medido em dólar ou em reais.

Um caso especial, atributos assimétricos.

- Representam dados (normalmente) binários onde apenas valores diferentes de zero são informativos.
- Por exemplo:
 - Dados de estudantes (linhas) x Disciplinas cursadas (colunas)

Quais são os tipos esses tipos?

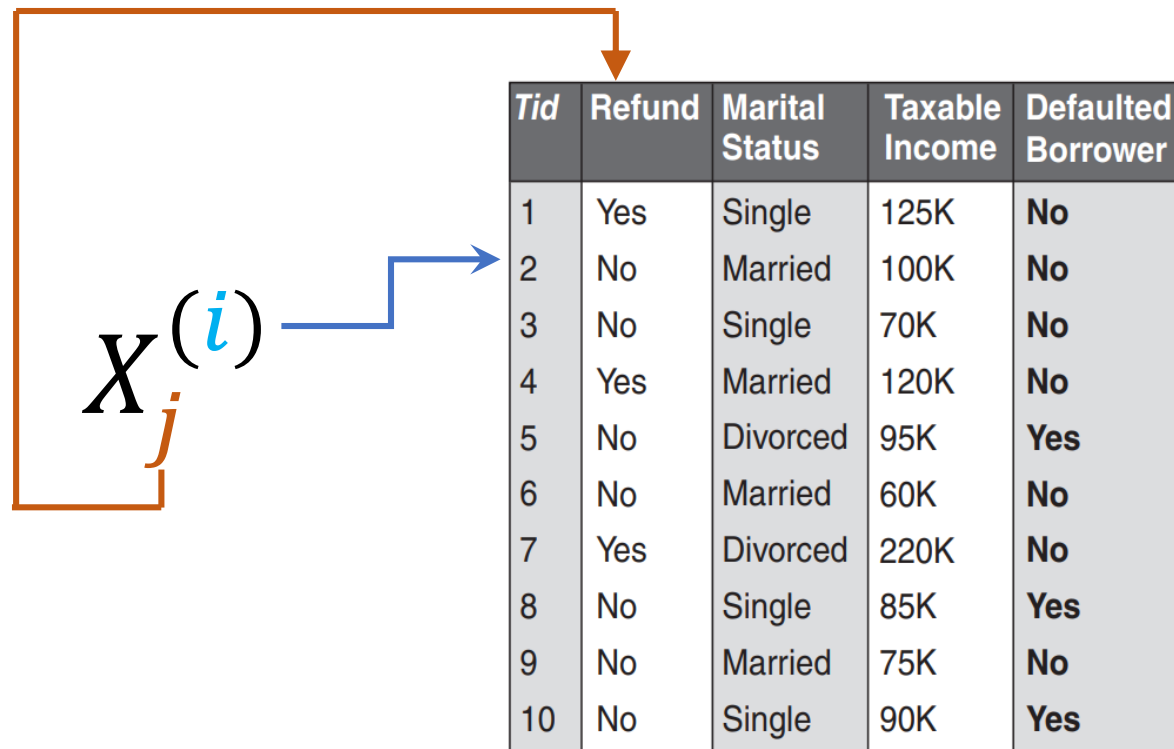
	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	class
0	25	Private	226802	11th	7	Never-married	Machine-op-inspct	Own-child	Black	Male	0	0	40	United-States	<=50K
1	38	Private	89814	HS-grad	9	Married-civ-spouse	Farming-fishing	Husband	White	Male	0	0	50	United-States	<=50K
2	28	Local-gov	336951	Assoc-acdm	12	Married-civ-spouse	Protective-serv	Husband	White	Male	0	0	40	United-States	>50K
3	44	Private	160323	Some-college	10	Married-civ-spouse	Machine-op-inspct	Husband	Black	Male	7688	0	40	United-States	>50K
4	18	?	103497	Some-college	10	Never-married	?	Own-child	White	Female	0	0	30	United-States	<=50K

Importante

- Capacidade de distinção, ordem, intervalos e razões são apenas quatro características dos dados.
- Outras formas de classificar dados existem (estruturados, cíclicos,...)
- O tipo de dado usado para armazenar uma *feature* pode levar você a cometer erros... Ex. computar média de Ids.
- Dados podem ser transformados para facilitar análises.

Conjuntos de dados baseados em “registros”

- A maioria do trabalho em análise de dados se baseia em dados ***tabulares*** ou na forma de ***registros***.



<i>Tid</i>	Refund	Marital Status	Taxable Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Tipos de dados tabulares

Dados de Transações

<i>TID</i>	<i>ITEMS</i>
1	Bread, Soda, Milk
2	Beer, Bread
3	Beer, Soda, Diapers, Milk
4	Beer, Bread, Diapers, Milk
5	Soda, Diapers, Milk

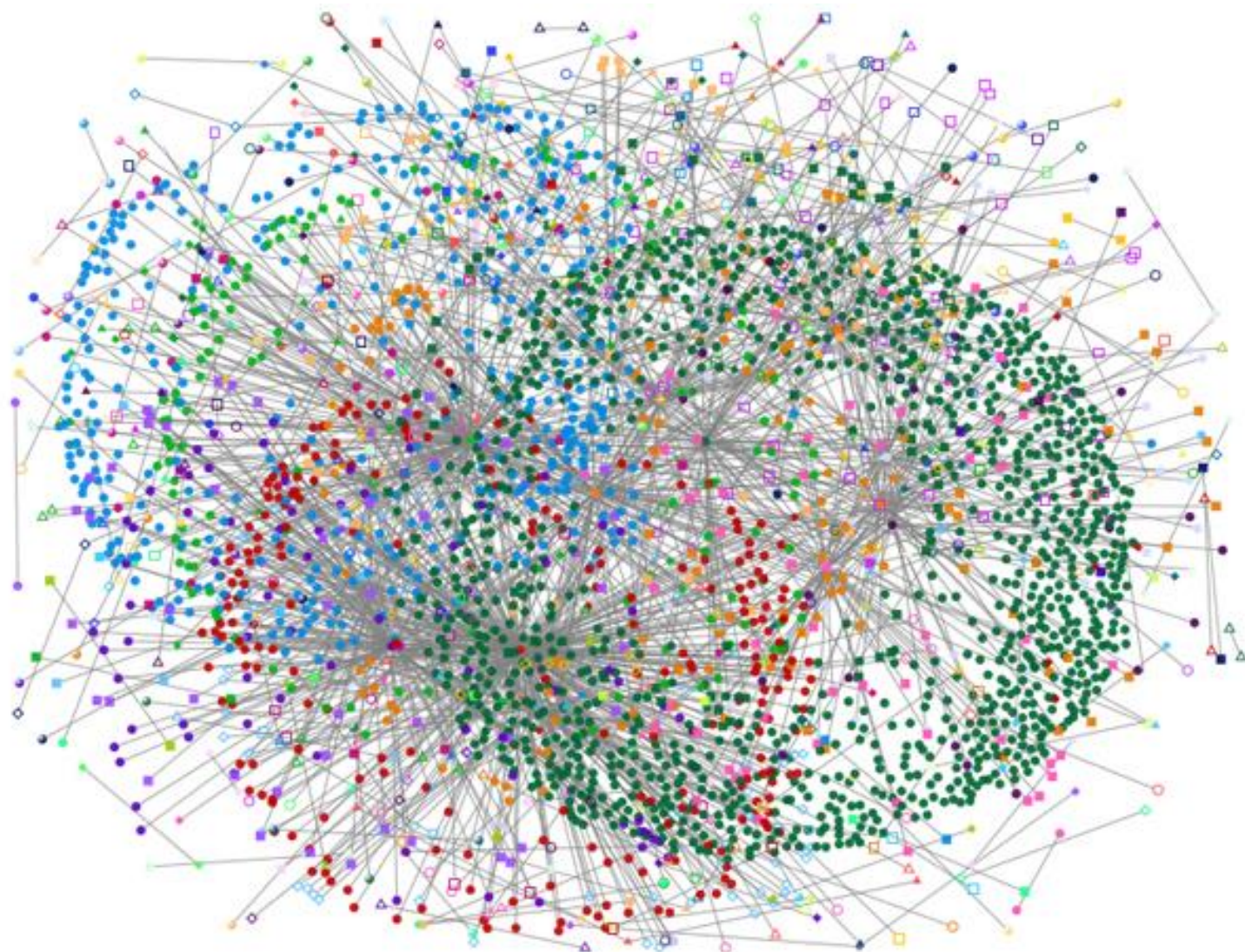
Matriz de Dados

Projection of x Load	Projection of y Load	Distance	Load	Thickness
10.23	5.27	15.22	27	1.2
12.65	6.25	16.22	22	1.1
13.54	7.23	17.34	23	1.2
14.27	8.43	18.45	25	0.9

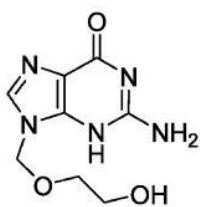
Matriz Termo-Documento

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

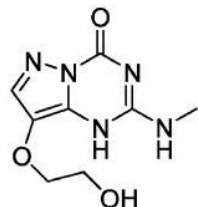
Dados em Grafos: Redes Sociais



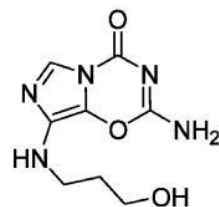
Dados em Grafos: Químicos



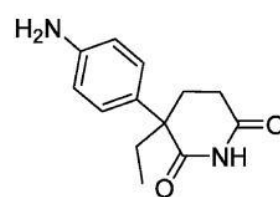
1 (Acyclovir)



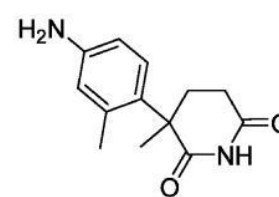
2



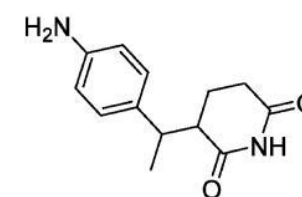
3



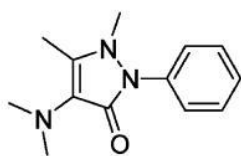
4 (Aminoglutethimide)



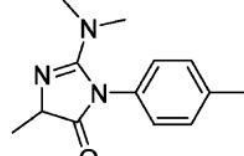
5



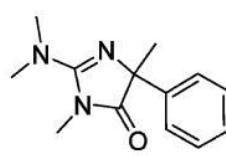
6



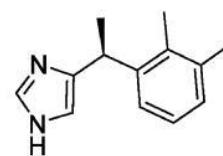
7 (Aminophenazone)



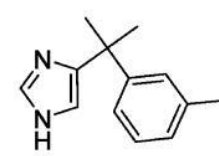
8



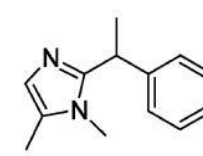
9



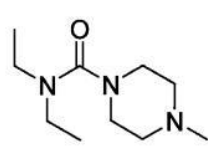
10 (Dexmedetomidine)



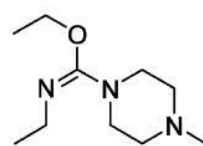
11



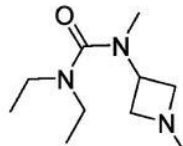
12



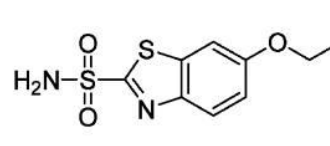
13 (Diethylcarbamazine)



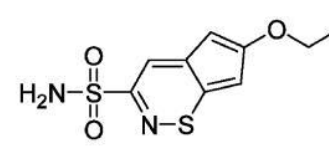
14



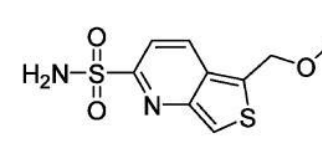
15



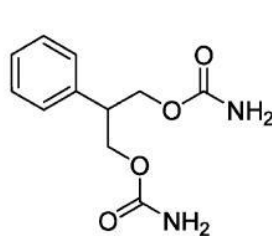
16 (Ethoxzolamide)



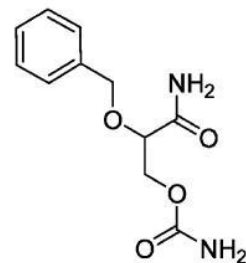
17



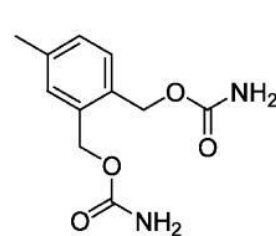
18



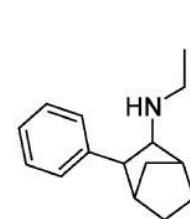
19 (Felbamate)



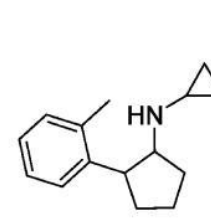
20



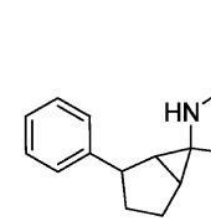
21



22 (Fencamfamine)

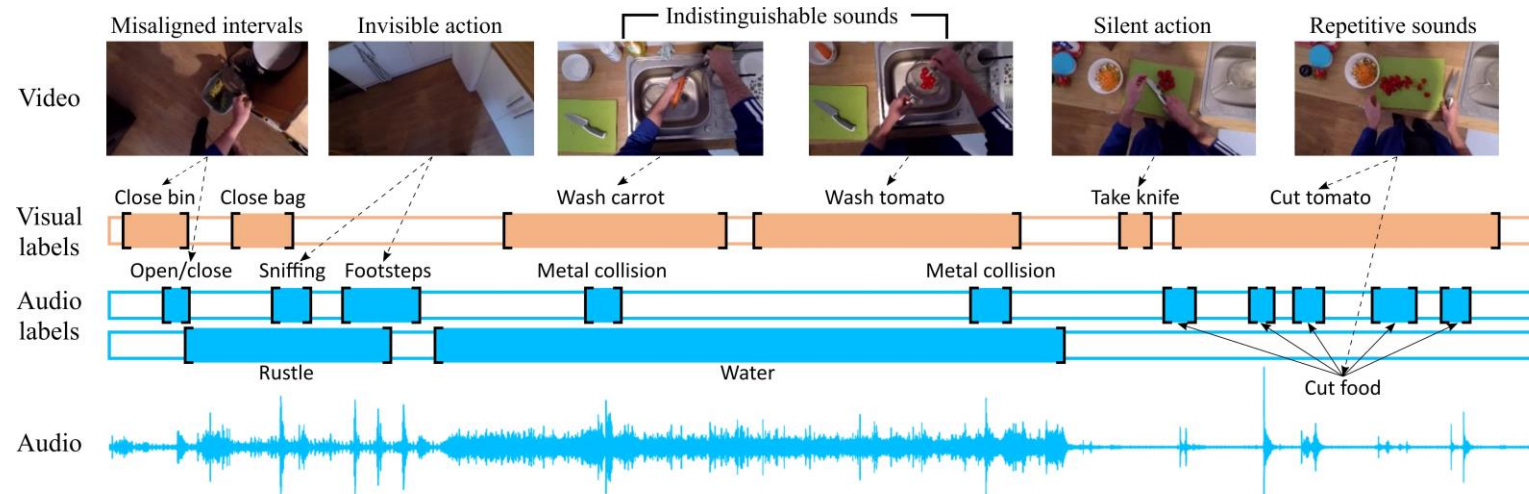
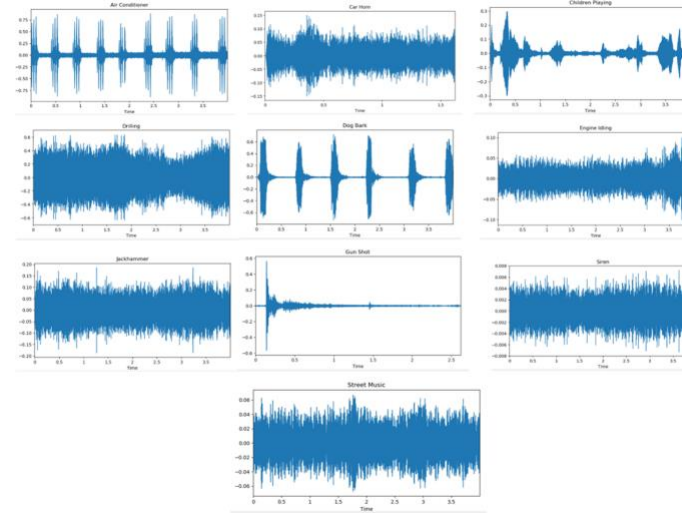


23



24

Dados Textuais, Imagens, Som e Vídeo

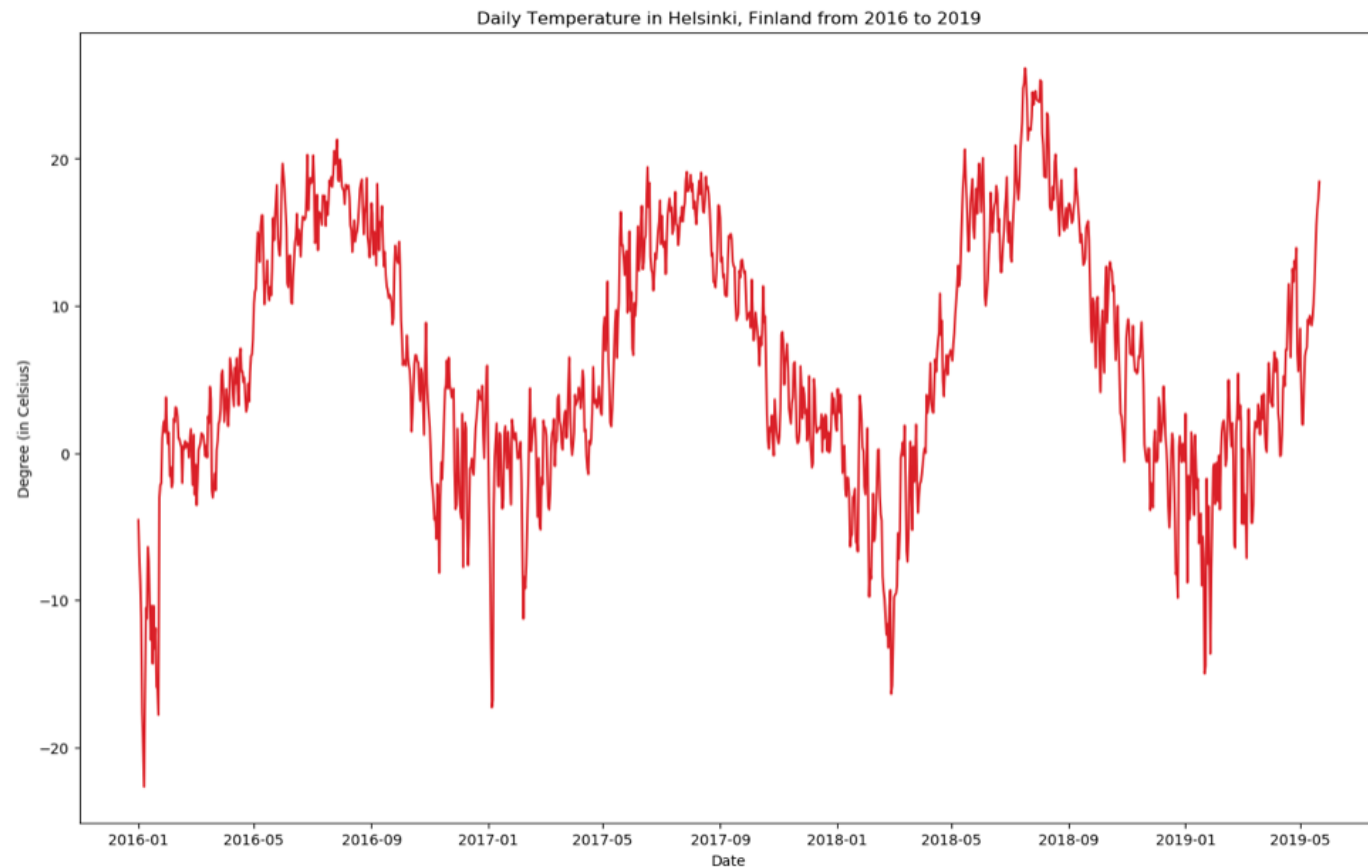


Dados Ordenados: Dados Genéticos

GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCCGCCCCGCGCCGTC
GAGAAGGGCCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG

Dados Ordenados – Séries Temporais

```
In [8]: # Plot the daily temperature change
plt.figure(figsize=(16,10), dpi=100)
plt.plot(temp_df.index, temp_df.T_mu, color='tab:red')
plt.gca().set(title="Daily Temperature in Helsinki, Finland from 2016 to 2019", xlabel='Date', ylabel="Degree (in Ce
plt.show()
```



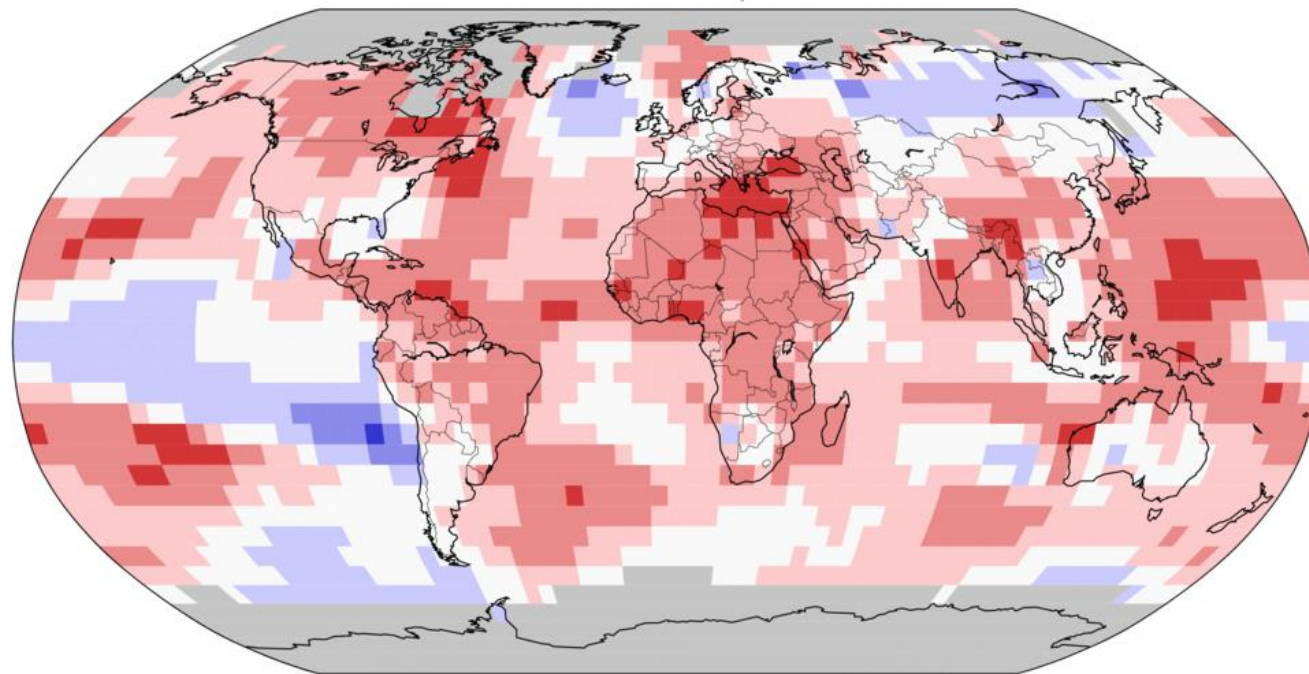
Dados Ordenados:

Dados de temperatura com informações posicionais.

Land & Ocean Temperature Percentiles Jan 2021

NOAA's National Centers for Environmental Information

Data Source: NOAA GlobalTemp v5.0.0-20210208



**Record
Coldest**



**Much
Cooler than
Average**



**Cooler than
Average**



**Near
Average**



**Warmer than
Average**



**Much
Warmer than
Average**



**Record
Warmest**

GHCNM v4.0.1.20210207.qfe

Dados ordenados - Desafios

- Dados “próximos” conforme a ordenação tendem a ser similares
- Dificuldade para análises preditivas e prescritivas.


Qualidade de Dados

Qualidade dos Dados

- **Corretude:** Dados precisam ser sempre corretos (livres de manipulação)
- **Confiabilidade:** Fonte dos dados deve ser confiável
- **Consistência:** Dados não podem estar corrompidos ou faltando
- **Atualidade:** Dados devem estar atualizados
- Exemplos de problemas na qualidade dos dados:
 - Ruídos e Outliers
 - Dados Faltantes
 - Dados Duplicados

Ruído

- O ruído ocorre quando há uma modificação nos valores originais dos dados
 - Exemplos contraditórios
 - Classificações incorretas
 - Valores incorretos para atributos



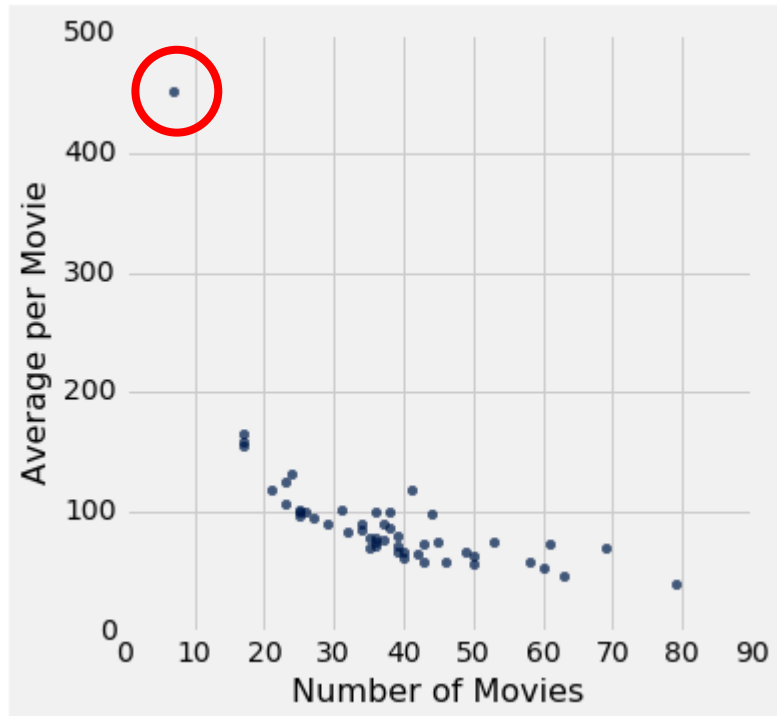
Atrib. 1	Atrib. 2	Classe
0.25	Vermelho	+
0.25	Vermelho	-
0.99	Verde	-
1.02	Verde	+

Valor Incorreto

Classificação Incorreta

Outliers

- Outliers são registros que possuem valores consideravelmente diferentes do que o resto dos outros objetos do dataset



Anthony Daniels

Valores Faltantes

- Múltiplas razões:
 - A informação não foi coletada (pessoa não quis responder sua idade e peso)
 - Atributos podem não ser aplicados em algum caso (salário não é aplicável a uma criança)
 - Desleixo (funcionário deixou de preencher campos na hora de fazer o cadastro)

Missing value

	loan_amnt	term	int_rate	sub_grade	emp_length	home_ownership	annual_inc	loan_status	addr_state	dti	mths_since_recent_inq	revol_util	bc_open_to_buy	bc_util	num_op_rev_tl
0	3600	36 months	14	C4	10+ years	MORTGAGE	55000	Fully Paid	PA	6	4	30	1506	37	4
1	24700	36 months	12	C1	10+ years	MORTGAGE	65000	Fully Paid	SD		0	19	57830	27	20
2	20000	60 months	11	B4	10+ years	MORTGAGE	63000	Fully Paid	IL		10	36	2737	56	4
3	35000	60 months	15	C5	10+ years	MORTGAGE		Current	NJ			12	54962	12	10
4	10400		12	F1	3 years	MORTGAGE	104433	Fully Paid	PA		1	64	4567	78	7
5			13	C3	4 years	RENT	34000	Fully Paid	GA	10		68	844	91	4
6	20000	36 months	9	B2	10+ years	MORTGAGE		Fully Paid	MN	15	10	84		103	9
7	20000	36 months	8	B1	10+ years	MORTGAGE	85000	Fully Paid	SC	18	8	6	13674	6	3
8		36 months	6	A2	6 years	RENT	85000	Fully Paid	PA	13	1	34		50	13
9		36 months	11	B5	10+ years	MORTGAGE	42000	Fully Paid	RI	35	10	39	9966	41	5

Dados Duplicados

- O dataset pode incluir registros que estão duplicados, ou quase duplicados
 - Esse problema acontece geralmente quando juntamos dados de diversas fontes

Nome do Professor
Lucas Rafael Costella Pessutto
Lucas R. C. Pessutto
Lucas Pessutto
Pessutto, Lucas C.

Pré-Processamento de Dados

Etapas de Pré-Processamento

- Agregação
- Amostragem
- Redução de Dimensionalidade
- Seleção de Features
- Criação de Features
- Transformação de Atributos

Agregação

- Combinação de dois ou mais registros em um único
- *Redução de dados*
 - Diminui o tamanho do dataset
- *Mudança de escala*
 - Cidades agregadas em estados agregados em países
- *Dados mais estáveis*
 - Dados agregados tendem a ter menos variabilidade

The diagram shows a transformation from a multi-year, multi-quarter dataset to a single-year, single-cost dataset. On the left, there are three overlapping tables for the years 2018, 2019, and 2020. The 2018 table is the most visible and contains four rows of quarterly data. An arrow points from this structure to a single table on the right that shows the aggregated annual costs for each year.

Year 2020	
Year 2019	
Year 2018	
Quater	Cost
Q1	\$224,000
Q2	\$408,000
Q3	\$350,000
Q4	\$586,000

Year	Cost
2018	\$1,568,000
2019	\$2,356,000
2020	\$3,594,000

Amostragem

- Técnica utilizada para realizar seleção de um conjunto menor de registros de um dataset
 - Muito usada para realizar análises preliminares e posteriores no conjunto de dados
- Amostragem também é utilizada quando processar todo o dataset demanda muito processamento e leva muito tempo
- Deve ser feita com cautela, para não descaracterizar os dados

Amostragem



8000 points



2000 Points



500 Points

Redução de Dimensionalidade

- Objetivo: reduzir os dados para poucas dimensões.
 - Maldição da dimensionalidade: Quando o número de dimensões aumenta, os dados tendem a ficar cada vez mais espalhados pelo espaço em que eles estão dispostos
- Porque?
 - Evitar a maldição da dimensionalidade
 - Reduzir tempo e memória necessários para executar algoritmos de aprendizagem de máquina
 - Permitir que os dados sejam visualizados mais facilmente
 - Identificação de features irrelevantes ou ruído
- Técnicas utilizadas:
 - Principal Component Analysis (PCA), Singular Value Decomposition (SVD), etc.

Seleção de Features

- Outra forma de reduzir a dimensionalidade dos dados
- Remoção de colunas redundantes
 - Informação duplicada ou muito parecida (cidade, UF, país)
- Features Irrelevantes
 - A informação não é útil para o problema
 - Ex: Número do RG e CPF de uma pessoa

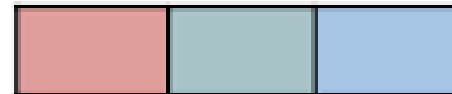
All Features



Feature Selection



Final Features



Criação de Features

- Criação de novas colunas no dataset, que capturem informações importantes dos dados mais eficientemente do que as colunas originais

- Técnicas utilizadas:

- Discretização
- One-hot-encoding
- Binarização
- Splitting
- Calculated features

Valor Categórico	Valor Inteiro	Valores Binarizados			One-Hot Encoding				
		x_1	x_2	x_3	x_1	x_2	x_3	x_4	x_5
Péssimo	0	0	0	0	0	0	0	0	1
Ruim	1	0	0	1	0	0	0	1	0
Regular	2	0	1	0	0	0	1	0	0
Bom	3	0	1	1	0	1	0	0	0
Ótimo	4	1	0	0	1	0	0	0	0

Transformação de Atributos

- Usar uma função que mapeia todo o conjunto de valores de um dado atributo para um novo conjunto de valores
 - Padronização e Normalização
- Exemplos:
 - Escala Logarítmica ($\log(x)$)
 - Normalização de case: upper / lower case