

Aprendizado de máquina

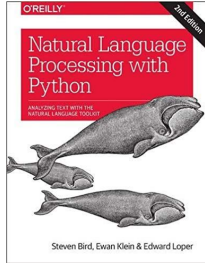
Fundamentos e aplicações em processamento de linguagem natural

Felipe Navarro Balbino Alves

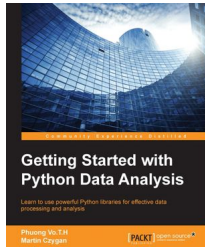
Github do curso

https://github.com/fnbalves/curso_machine_learning/

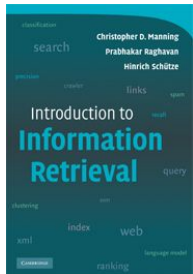
Bibliografia de interesse



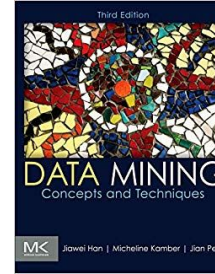
Natural Language Processing with Python
Steven Bird (Disponível online)



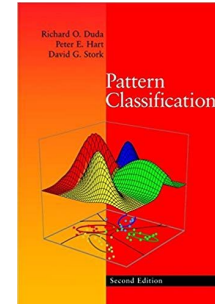
Getting started with Python Data Analysis
Phuong Vo. T.H



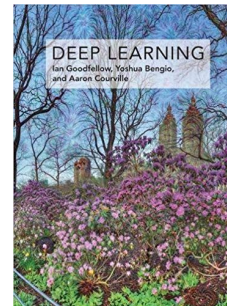
Introduction to information Retrieval
Cristopher D. Manning



Data Mining - Concepts and Techniques
Han & Kamber



Pattern Classification
Han & Kamber



Deep Learning
Ian Goodfellow (Disponível online)

Bibliografia de interesse - Material online

NLTK book:

<https://www.nltk.org/book/>

Introduction to Information Retrieval:

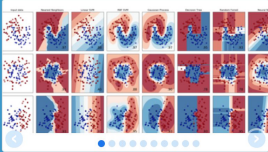
<https://nlp.stanford.edu/IR-book/>

Scikit-learn documentation:

<https://scikit-learn.org/stable/>

The Deep Learning book:

<https://www.deeplearningbook.org/>



scikit-learn

Machine Learning in Python

- Simple and efficient tools for data mining and data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

<h3>Classification</h3> <p>Identifying to which category an object belongs to.</p> <p>Applications: Spam detection, Image recognition.</p> <p>Algorithms: SVM, nearest neighbors, random forest, ... — Examples</p>	<h3>Regression</h3> <p>Predicting a continuous-valued attribute associated with an object.</p> <p>Applications: Drug response, Stock prices.</p> <p>Algorithms: SVR, ridge regression, Lasso, ... — Examples</p>	<h3>Clustering</h3> <p>Automatic grouping of similar objects into sets.</p> <p>Applications: Customer segmentation, Grouping experiment outcomes</p> <p>Algorithms: k-Means, spectral clustering, mean-shift, ... — Examples</p>
<h3>Dimensionality reduction</h3> <p>Reducing the number of random variables to consider.</p> <p>Applications: Visualization, Increased efficiency</p> <p>Algorithms: PCA, feature selection, non-negative matrix factorization. — Examples</p>	<h3>Model selection</h3> <p>Comparing, validating and choosing parameters and models.</p> <p>Goal: Improved accuracy via parameter tuning</p> <p>Modules: grid search, cross validation, metrics. — Examples</p>	<h3>Preprocessing</h3> <p>Feature extraction and normalization.</p> <p>Application: Transforming input data such as text for use with machine learning algorithms.</p> <p>Modules: preprocessing, feature extraction. — Examples</p>

O que é aprendizado de máquina?

Exemplo prático: precificar um imóvel

Seu trabalho é ajudar um corretor de imóveis a estimar o preço de uma propriedade.

Como podemos atacar o problema?

Valor Venda		R\$ 290.000		
				
Quartos	Banheiros	M² total	Vagas	
2	2	120	1	

O que é aprendizado de máquina?

Exemplo prático: precificar um imóvel

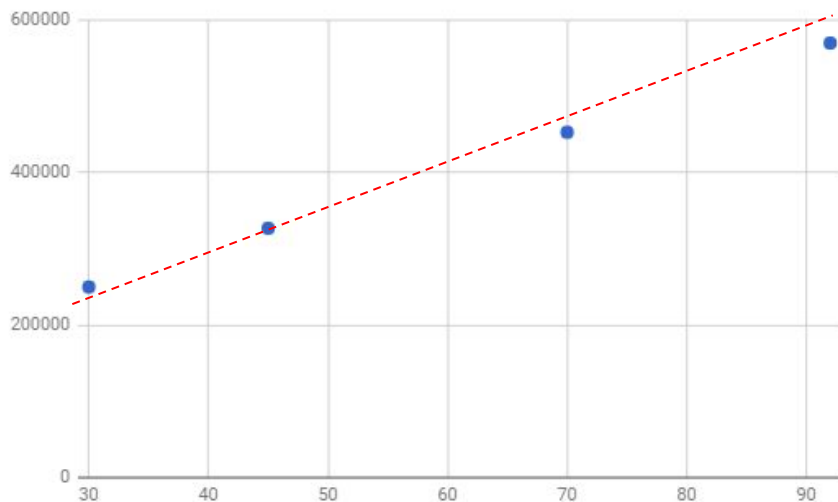
O corretor lhe fornece a seguinte planilha de dados de precificação anteriores

Imóvel	Tamanho do imóvel em metros quadrados	Preço do imóvel em reais
Rua da Palma, 467	30	250.000
Conde de irajá 344	45	325.100
Agamenon magalhães 56	70	453.000
Joaquim Nabuco 443	92	570.000

O que é aprendizado de máquina?

Exemplo prático: precificar um imóvel

Você decide plotar os dados



Você descobre que a relação é aproximadamente uma reta, de equação

$$100000 + 5000 \times (\text{tamanho do imóvel})$$

O que é aprendizado de máquina?

Exemplo prático: precificar um imóvel

Agora vamos supor que a planilha fornecida fosse assim:

Imóvel	Tamanho do imóvel (m ²)	Número de quartos	Tem gás encanado	Impostos em dia?	Vagas de garagem	Cidade	Construtora
Imóvel 1	30	1	Sim	Não	1	Recife	(dado faltante)
Imóvel 2	45	3	Não	Sim	2	Jaboatão	Construtora legal
Imóvel 3	70	2	Não	(dado faltante)	2	Paulista	A sua construtora
Imóvel 4	92	5	Sim	Não	3	Olinda	A melhor construtora

...

O que é aprendizado de máquina?

Exemplo prático: precificar um imóvel

E agora, como fazemos para **extrair uma regra** de forma visual?

Imóvel	Tamanho do imóvel (m ²)	Número de quartos	Tem gás encanado	Impostos em dia?	Vagas de garagem	Cidade	Construtora
Imóvel 1	30	1	Sim	Não	1	Recife	(dado faltante)
Imóvel 2	45	3	Não	Sim	2	Jaboatão	Construtora legal
Imóvel 3	70	2	Não	(dado faltante)	2	Paulista	A sua construtora
Imóvel 4	92	5	Sim	Não	3	Olinda	A melhor construtora

...

O que é aprendizado de máquina?

Aprendizado de máquina é um sub-ramo da ciência da computação especializado no reconhecimento automático de **padrões** a partir de **dados**

Inteligência artificial x aprendizado de máquina

Inteligência artificial é um conceito mais amplo e trata de máquinas capazes de realizar tarefas consideradas “inteligentes”. Abrange temas como Teoria dos jogos, Sistemas de busca, representação de conhecimento, planejamento, entre outros

Inteligência artificial x aprendizado de máquina

● machine learning

Termo de pesquisa

● artificial intelligence

Termo de pesquisa

+ Adicionar comparação

Estados Unidos ▼

Nos últimos 5 anos ▼

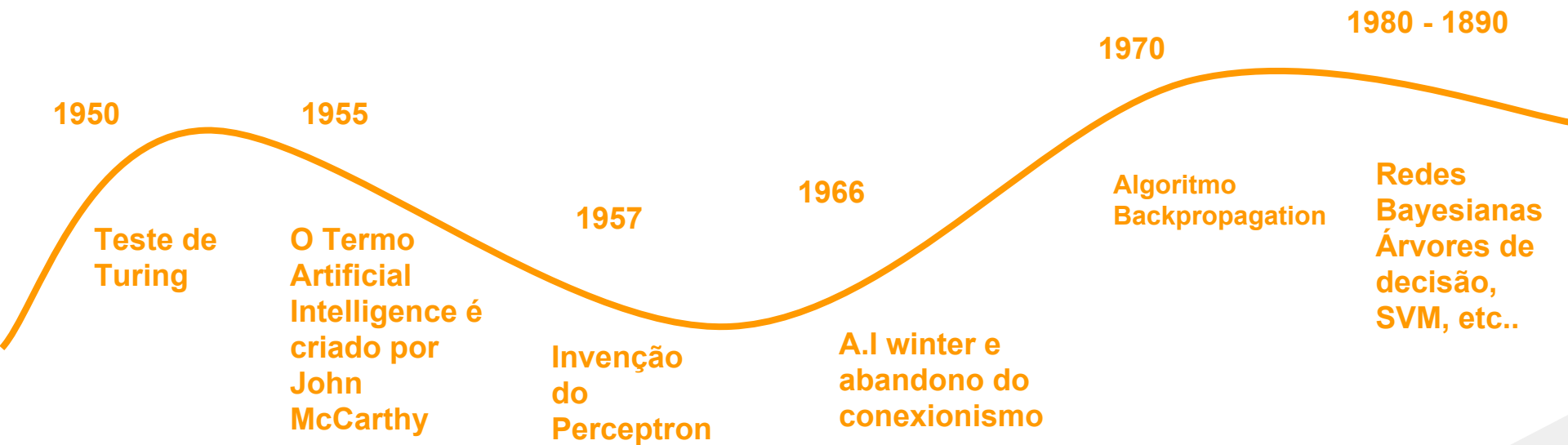
Todas as categorias ▼

Pesquisa na Web ▼

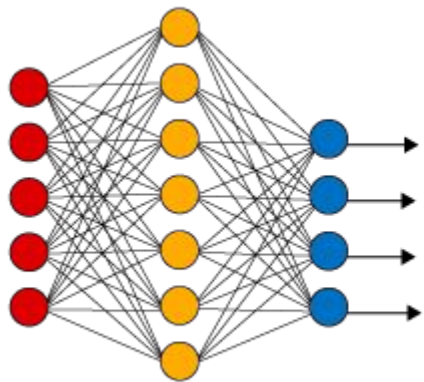
Interesse ao longo do tempo ?



Evolução do AM

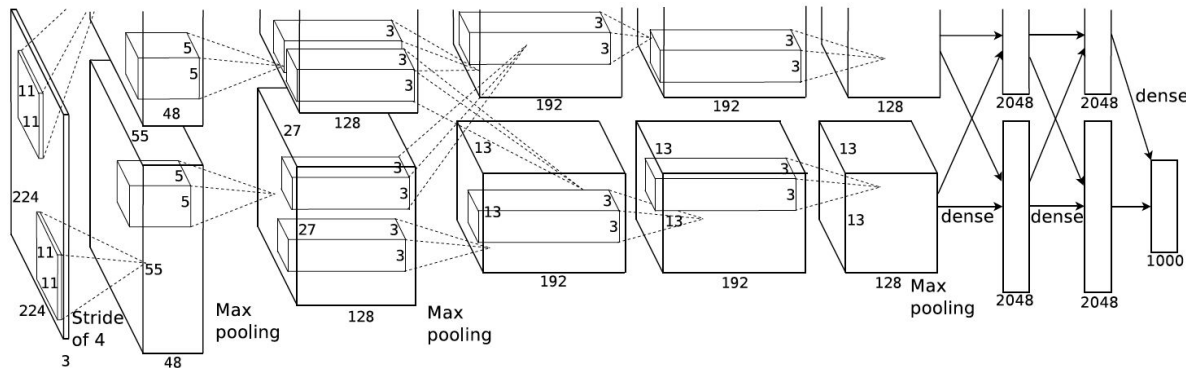


Surgimento do deep learning

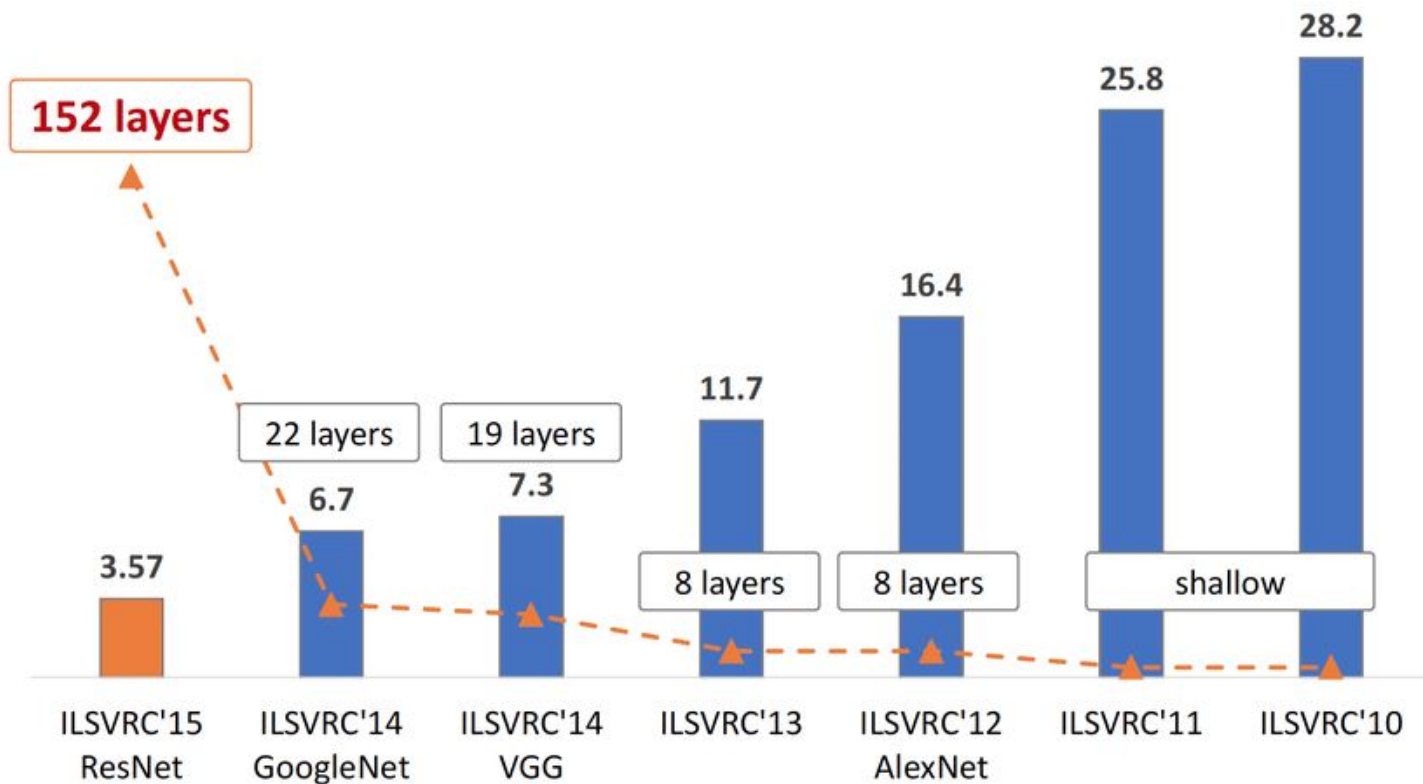


Shallow network

Vs Alex net



Surgimento do deep learning



Evolução de poder computacional

GOOGLE BRAIN

1,000 CPU Servers
2,000 CPUs • 16,000 cores

600 kWatts
\$5,000,000

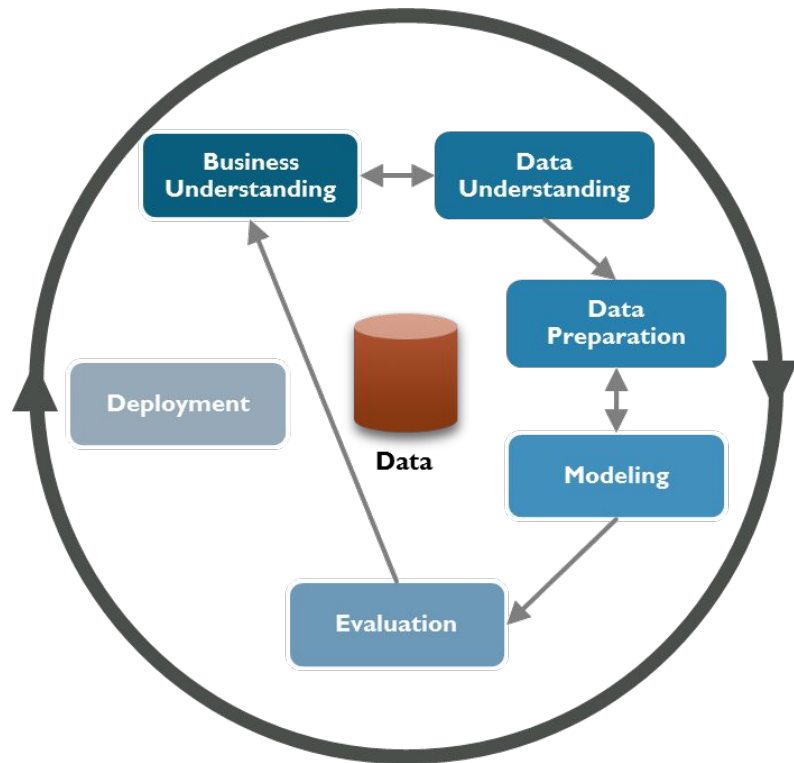
300X energy efficiency
400X lower cost
Fits under a desk



1 Titan Z-Accelerated Server
3 Titan Zs • 17,280 cores

2 kWatts
\$12,000

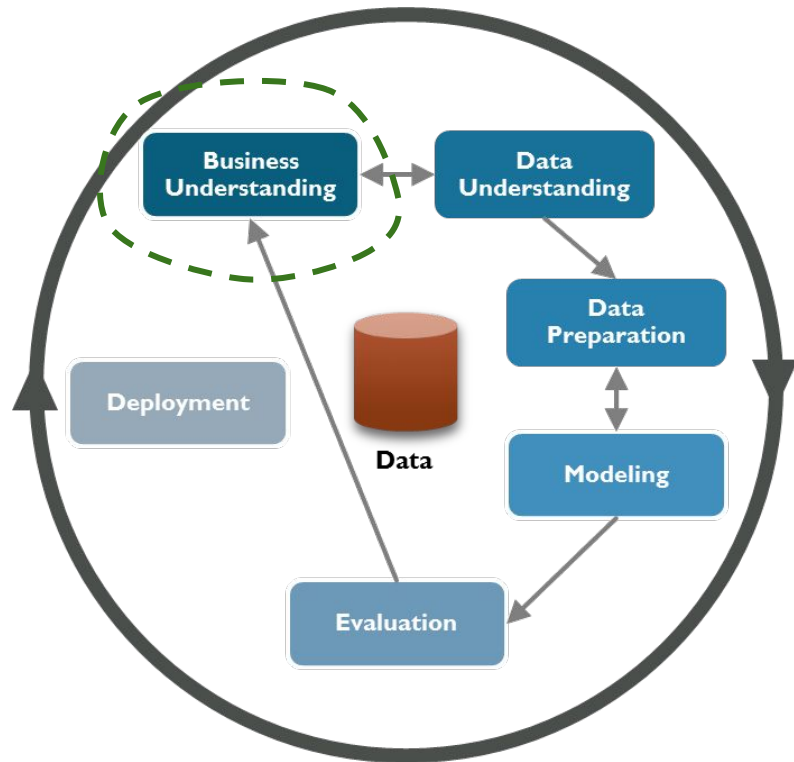
Como funciona um projeto de Aprendizado de máquina



CRISP - DM

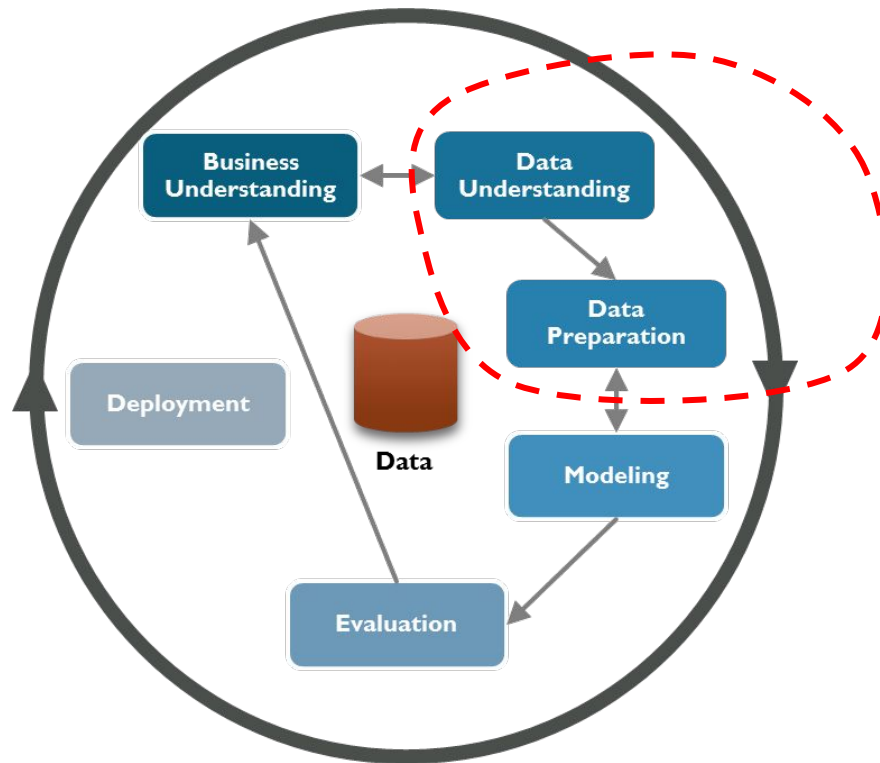
Como funciona um projeto de Aprendizado de máquina

Identificação
de
oportunidade



CRISP - DM

Como funciona um projeto de Aprendizado de máquina



Entendimento do significado das variáveis

Caracterização estatística das variáveis

Limpeza e tratamento da base de dados

CRISP - DM

Em relação aos dados

O primeiro passo para a correta utilização de uma variável é entender o que ela representa

Dessa forma, evitamos utilizar dados **a posteriori**

Exemplo: Suponha que se deseja criar um modelo para prever a complexidade do conserto de uma máquina a partir do log de eventos da mesma. Podemos usar como feature a **peça que foi substituída?**

Caracterização estatística dos dados

A seguir, precisamos identificar qual o tipo da variável em questão:

Variável numérica: Tamanho do terreno (30m², 49 m²,)

Variável categórica: IPTU em dia? (Sim ou não)

Dentro das variáveis categóricas, podemos classificá-las em:

Variável nominal: Não existe uma ordem de grandeza. Ex: sexo, estado civil

Variável ordinal: Existe uma ordem entre as categorias. Ex: escolaridade

Uma variável categórica pode ser nominal ou ordinal **dependendo do contexto.**

Caracterização estatística dos dados

Para **variáveis numéricas**, é importante identificar durante o entendimento dos dados, as seguintes grandezas:

Porcentagem de Missing data

Média

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^n x_i$$

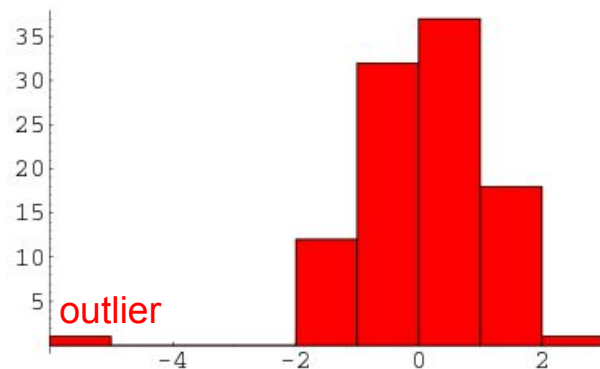
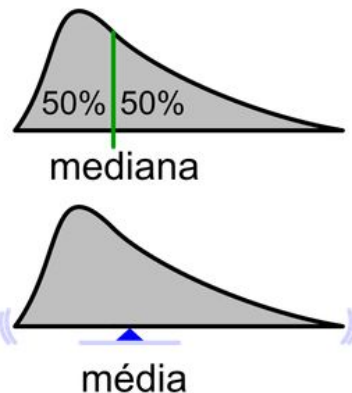
Variância

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

Outliers

Mediana

Histograma



Caracterização estatística dos dados

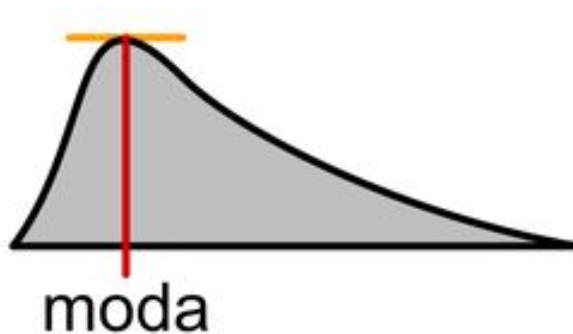
Para **variáveis categóricas**, é importante identificar durante o entendimento dos dados, as seguintes grandezas:

Porcentagem de Missing data

Moda

Outliers

Histograma



Análise de correlação

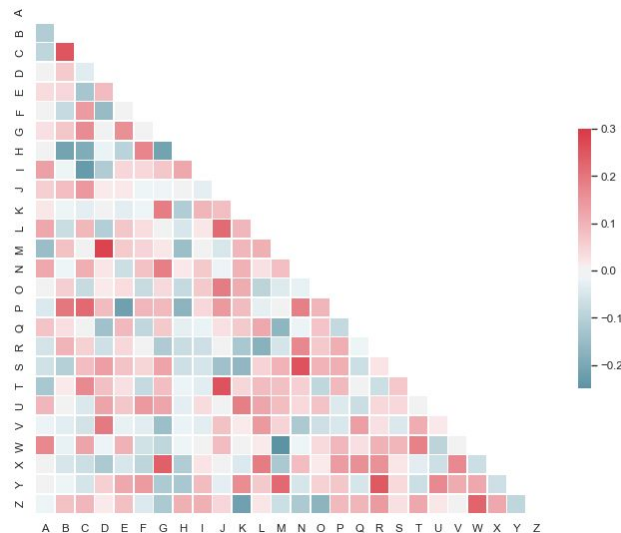
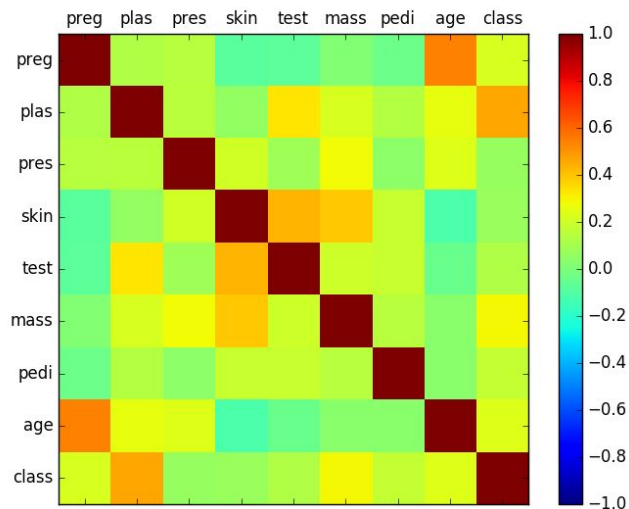
Muitas vezes, uma mesma informação é representada de múltiplas formas em uma mesma base de dados. Variáveis redundantes são um **problema** para muitos algoritmos de aprendizado. Desta forma, se faz necessária uma análise de correlação

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

Se $\rho > 0$, as variáveis tendem a crescer ou decrescer ao mesmo tempo. Se $\rho < 0$, as variáveis tendem a ter comportamento oposto.

Análise de correlação

Nas bibliotecas que mostraremos ao longo do treinamento, a análise de correlação pode ser feita a partir de uma **matriz de correlação**



Missing data

O tratamento de missing data varia bastante de acordo com o contexto e com o negócio. É necessário tentar entender qual o **motivo** do aparecimento do valor faltante. Para variáveis numéricas, temos algumas opções:

1. **Listwise deletion:** Eliminar registros com o missing data;
2. **Average imputation:** Substituir um dado faltante pela média dos demais;
3. **Regression substitution:** Criação de um modelo para prever o missing value.

Missing data

No caso de variáveis categóricas, as opções são um pouco diferentes:

1. **Listwise deletion:** Eliminar registros com o missing data;
2. **Average imputation:** Substituir um dado faltante pela **moda** dos demais;
3. **Classification substitution:** Criação de um modelo para prever o missing value;
4. Criação de uma categoria extra para identificar missing values;

Engenharia de atributos

Suponha que se deseje criar um modelo para calcular a probabilidade de um indivíduo ter uma **doença vascular**.

Suponha que você possua dois parâmetros: peso (kg) e altura (m).

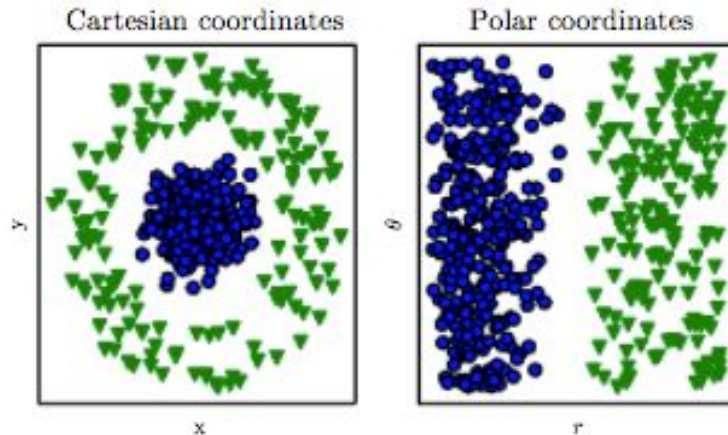
Provavelmente, estas duas variáveis em conjunto não serão mais relevantes do que a variável combinada:

$$\text{imc} = \text{peso} / \text{altura}^2$$

Engenharia de atributos

Para criar novas variáveis de **grande relevância**, é, em geral necessário um grande conhecimento específico do domínio.

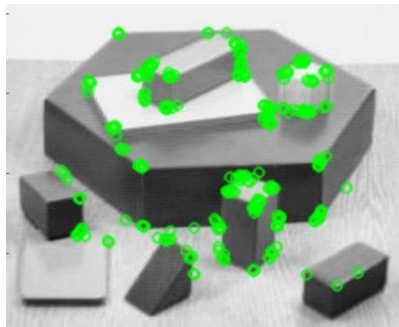
Exemplo 2:



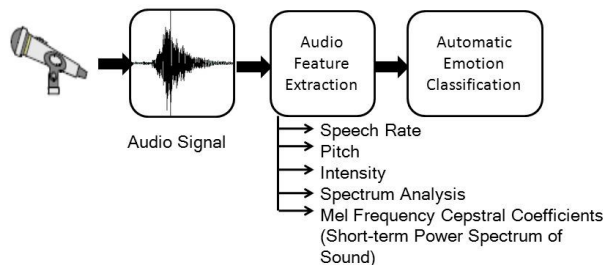
Engenharia de atributos

A engenharia de features é utilizada em todas as modalidades do aprendizado de máquina. Geralmente exigem **conversas com especialistas**

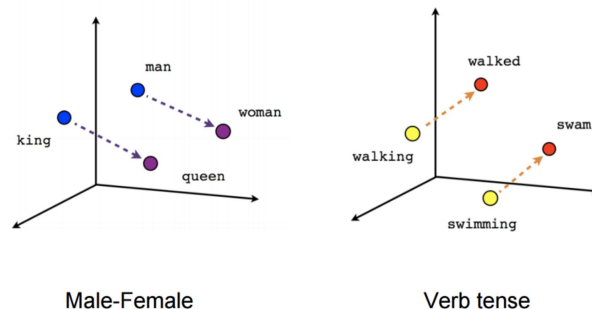
Processamento de imagens



Audio Feature Extraction



Processamento de texto



Mudança de granularidade

Com frequência o processo de análise de dados exige a mudança da granularidade da informação. Como exemplo, suponha que temos à disposição dados de **alunos** do ensino médio, mas estamos interessados em analisar **instituições** de ensino médio.

Como devemos lidar com variáveis como a idade dos alunos, a escolaridade dos pais, etnia, etc... quando analisamos a instituição como um todo?

Mudança de granularidade - variáveis numéricas

Para variáveis numéricas, podemos criar novas variáveis que representam características estatísticas dos dados

Média

Desvio padrão

Max

Min

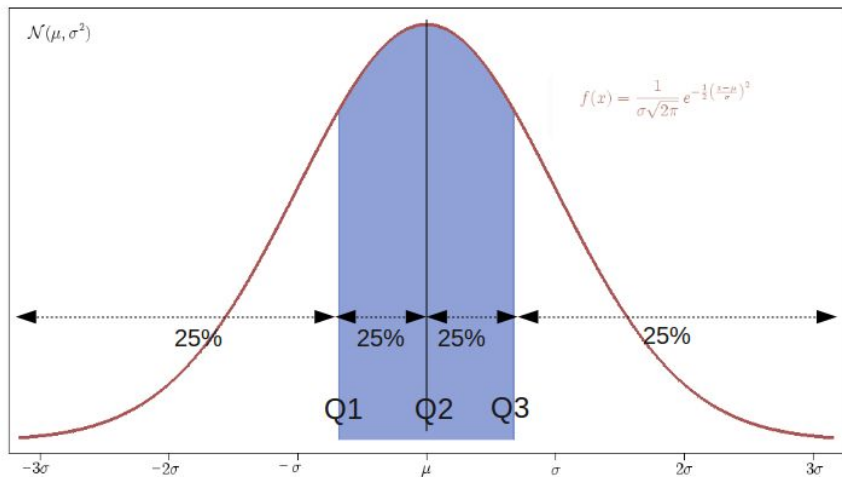
Mediana

Moda

Quantis

Quantis

Quantis são valores que dividem a área sob a função distribuição de probabilidade em partes iguais

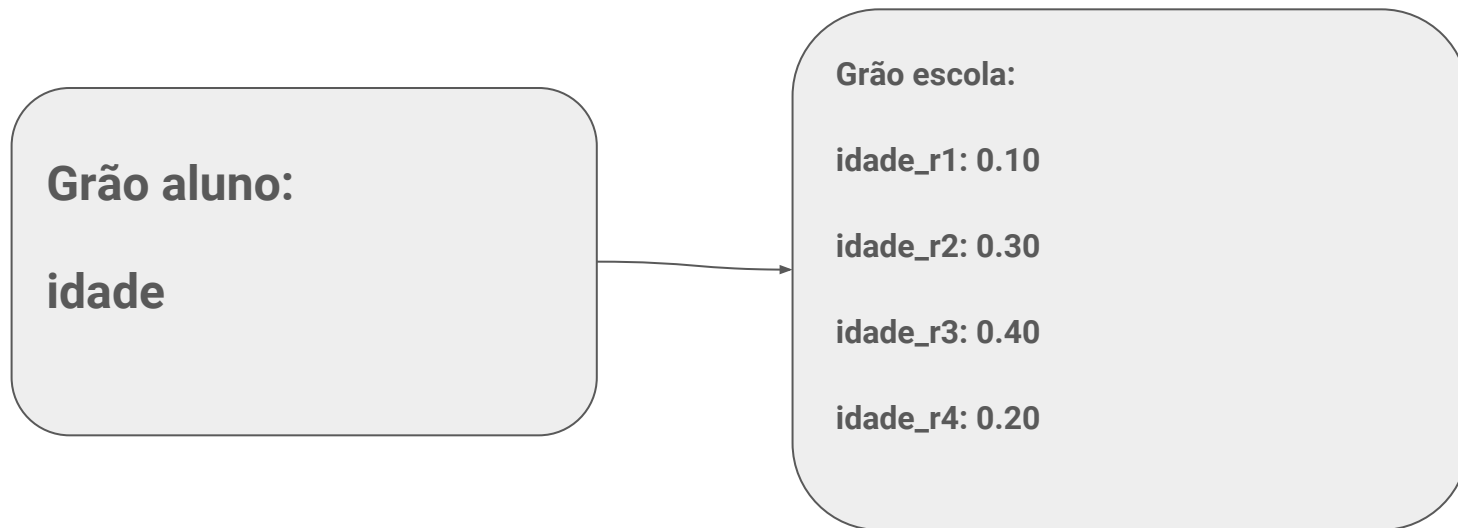


No gráfico maior, podemos colocar a quantidade de elementos que apareceu em cada uma das regiões

Quantis

Quantis são valores que dividem a área sob a função distribuição de probabilidade em partes iguais

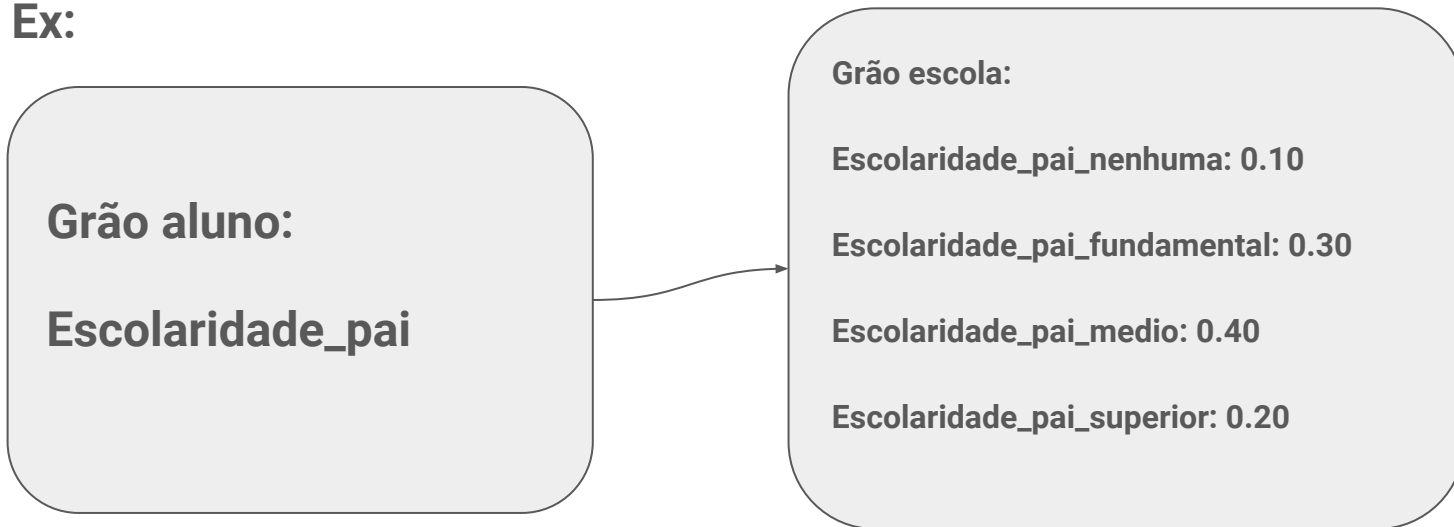
Ex:



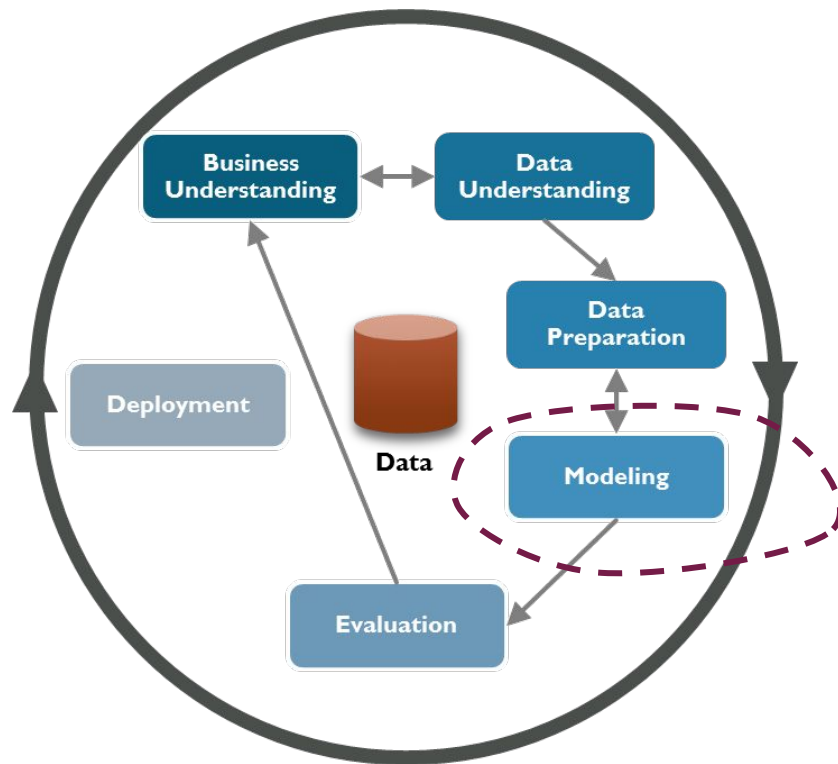
Mudança de granularidade - variáveis categóricas

Para variáveis categóricas, podemos fazer algo similar à estratégia dos quantis, inserindo quantos elementos apareceram em cada classe, ou até mesmo inserindo a frequência relativa

Ex:



Como funciona um projeto de Aprendizado de máquina



Neste momento, tentaremos identificar o padrão nos dados

CRISP - DM

Tipos de modelos

Podemos dividir os modelos de aprendizado de máquina em duas grandes classes: modelos supervisionados e modelos não supervisionados.

O exemplo de precificação de um imóvel é uma forma de **aprendizado supervisionado**, onde existe um “professor” que diz a resposta correta para vários exemplos de entrada. Dizemos que temos **dados rotulados**.

Existe o **aprendizado não supervisionado**, onde não existem dados rotulados, o objetivo é encontrar **grupos de similaridade**

Aprendizado supervisionado

Dentro do aprendizado supervisionado, podemos ter dois tipos de problemas

Regressão: Exemplo do preço da casa

Classificação: Classificar uma entrada em um número discreto de possibilidades
(Ex: reconhecer se uma imagem é um gato ou cachorro)

Ganho de informação

No caso de algoritmos de classificação, existe outra métrica que pode ser utilizada para analisar os dados, que é o **Ganho de informação**, que se baseia na medida **entropia**

$$Entropia(S) = - \sum p_i \log_2 p_i$$

Dois casos:

classe 1:
probabilidade (50%)

classe 2:
probabilidade (50%)

$$Entropia = -0.5 \cdot \log_2(0.5) + -0.5 \cdot \log_2(0.5) = 1$$

classe 1:
probabilidade (90%)

classe 2:
probabilidade (10%)

$$Entropia = -0.9 \cdot \log_2(0.9) + -0.1 \cdot \log_2(0.1) = 0.46$$

Ganho de informação

O Ganho de informação de uma variável é definida da seguinte forma:

$$IG(T, a) = H(T) - H(T|a).$$

$$S_a(v) = \{\mathbf{x} \in T | x_a = v\}$$

$$H(T|a) = \sum_{v \in \text{vals}(a)} \frac{|S_a(v)|}{|T|} \cdot H(S_a(v))$$

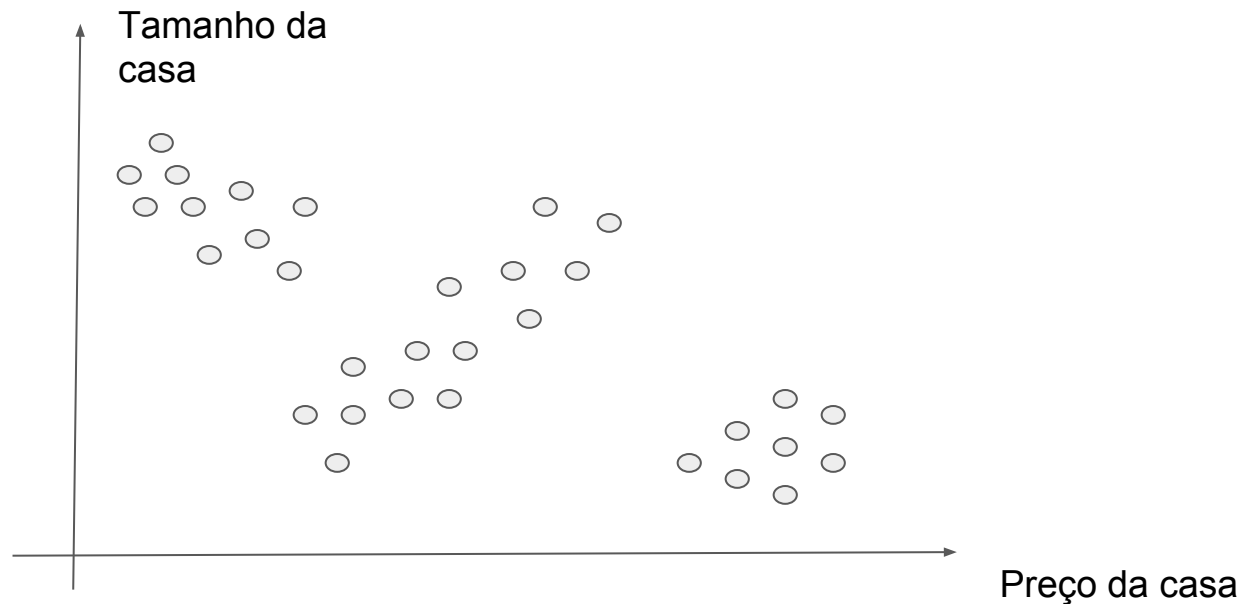
Variáveis com maior ganho de informação tem maior chance de serem relevantes no processo de classificação

Aprendizado não supervisionado

No **aprendizado não supervisionado** , onde não existem dados rotulados, o objetivo é encontrar **grupos de similaridade**

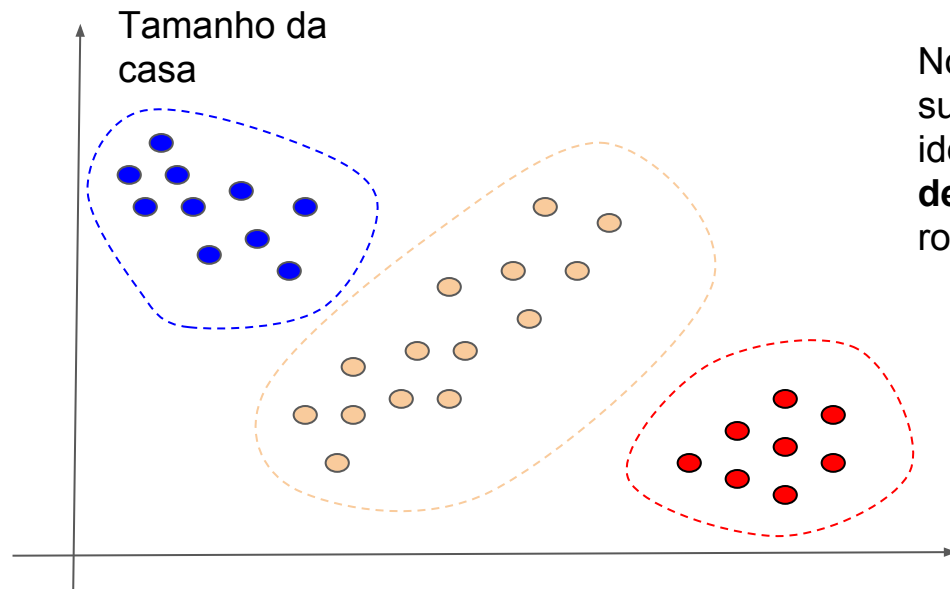
Aprendizado não supervisionado

Exemplo



Formas de aprendizado

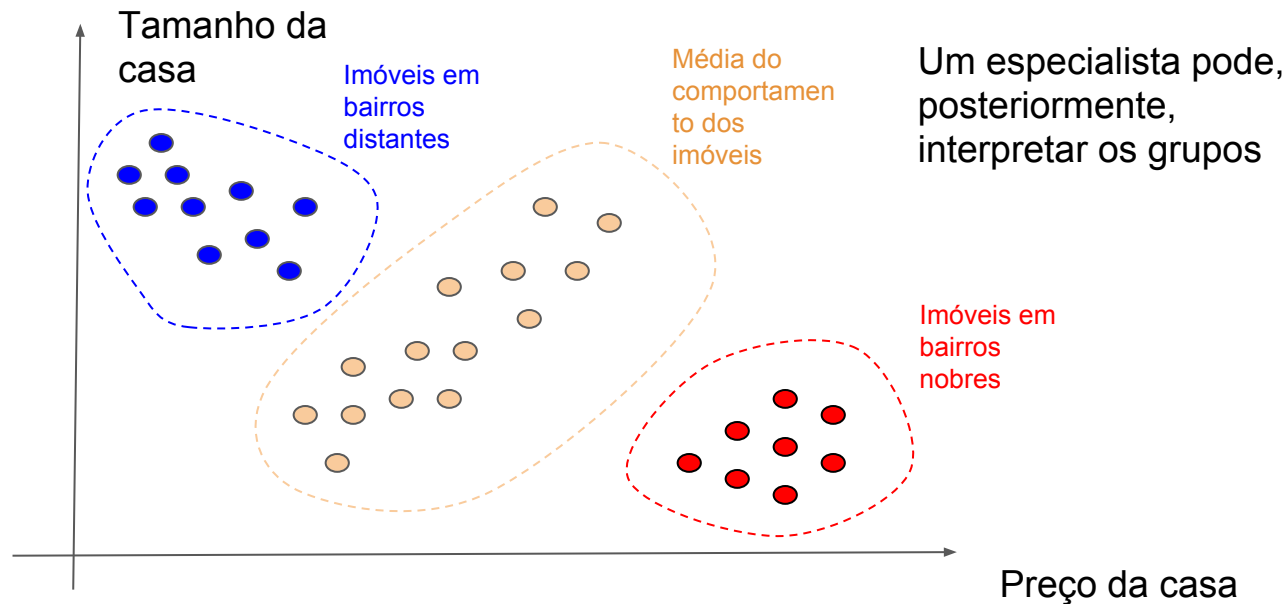
Exemplo



No aprendizado não supervisionado, são identificados **grupos de similaridade** não rotulados

Formas de aprendizado

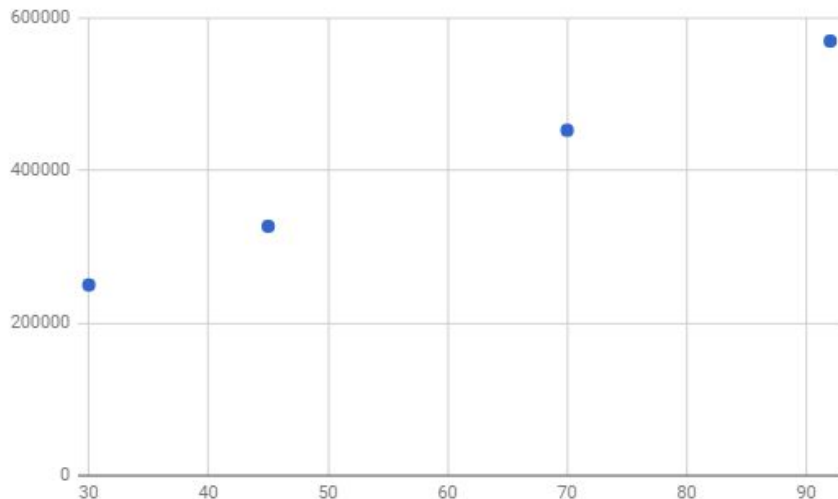
Exemplo



Analizando o padrão encontrado

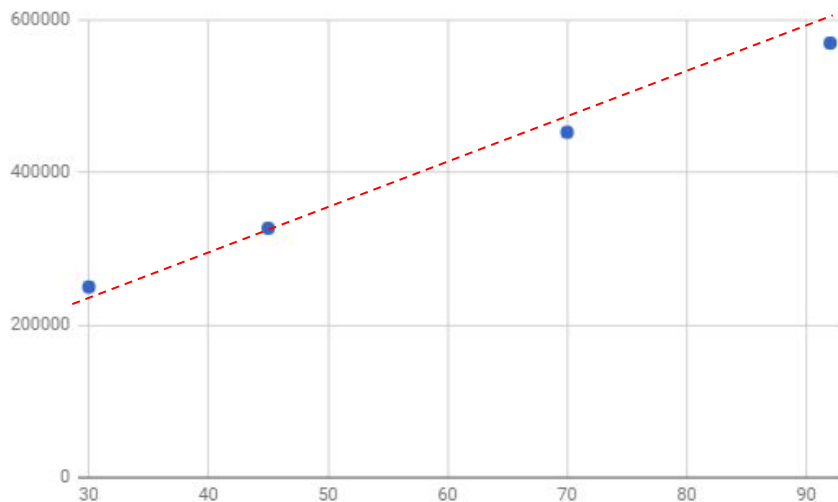
Exemplo prático: precificar um imóvel

Voltando ao exemplo da precificação do imóvel. Temos os dados. Qual é o **padrão**?



Avaliando o padrão encontrado

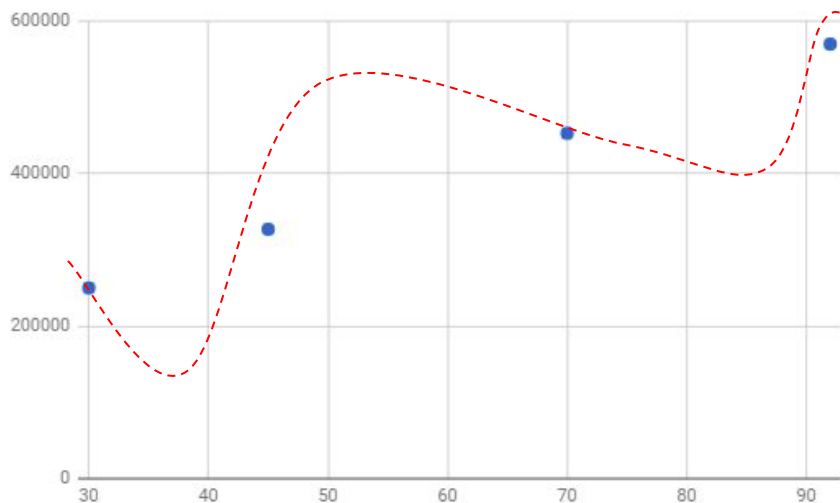
No âmbito da ciência de dados, o padrão também é chamado de **hipótese**



Neste caso, a nossa hipótese é que a lei que regia o fenômeno era uma linha reta, mas **precisava ser?**

Qual o padrão?

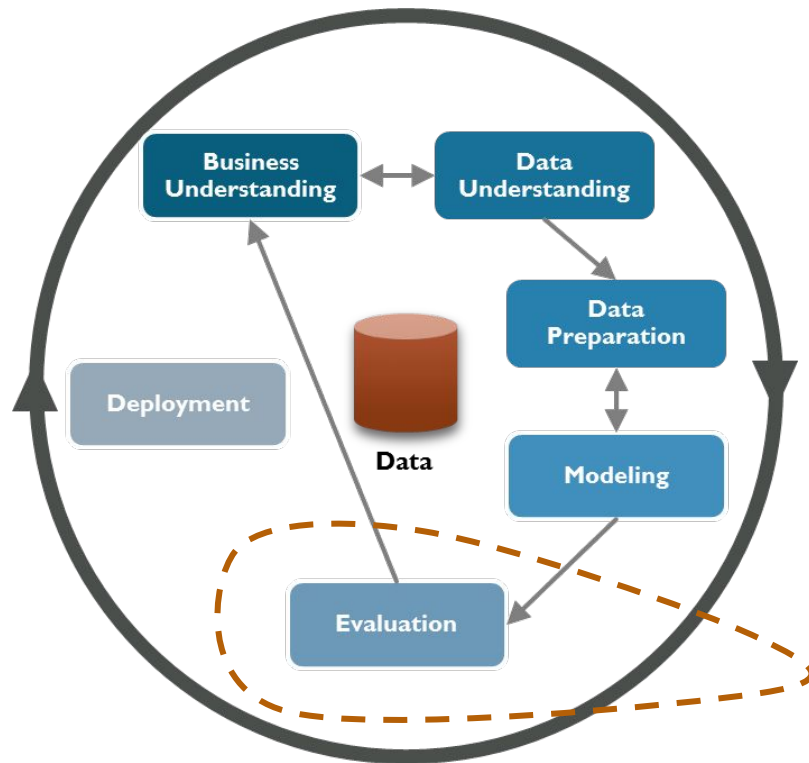
No âmbito da ciência de dados, o padrão também é chamado de **hipótese**



Esta outra hipótese também modela perfeitamente os **dados de treinamento**

Será que ela é melhor?

Como funciona um projeto de Aprendizado de máquina



O exemplo anterior mostra a necessidade de métricas de avaliação dos modelos

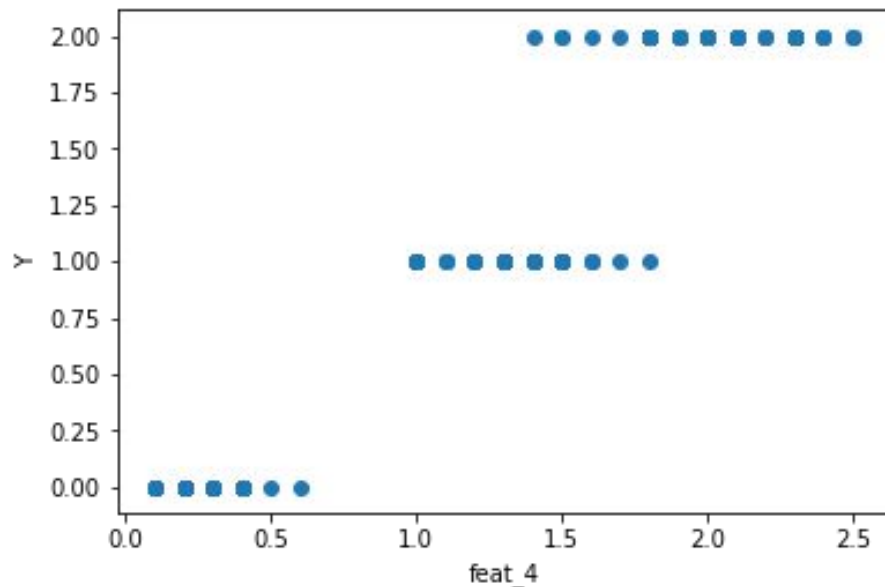
CRISP - DM

Visualização de dados

Como sugerido pelo exemplo da precificação de imóveis, uma forma do cientista de dados ter ideias a respeito de quais **hipóteses** testar é a **visualização dos dados**

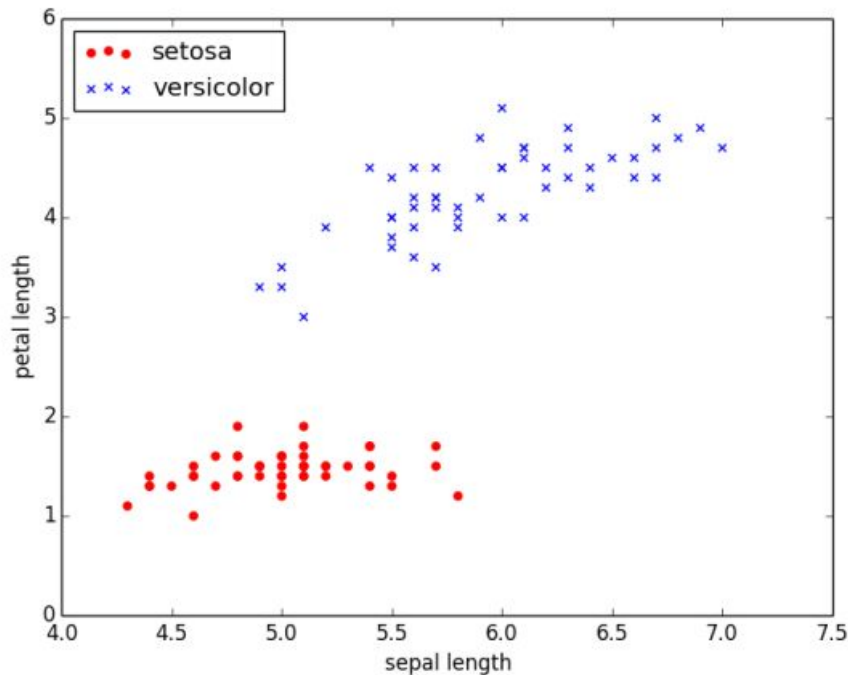
Visualização de dados

Uma possibilidade é dispor a relação entre **features individuais** e a **variável objetivo** em um gráfico de dispersão ou **scatter plot**



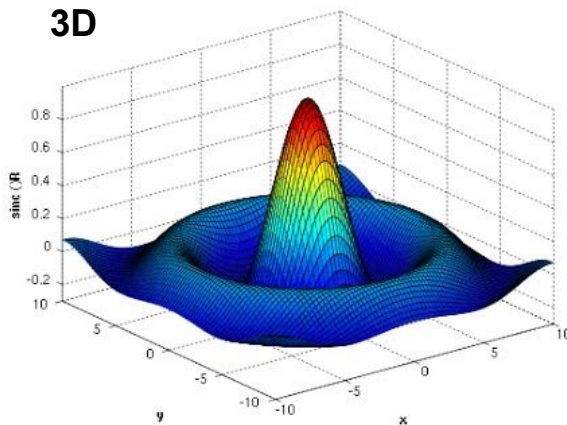
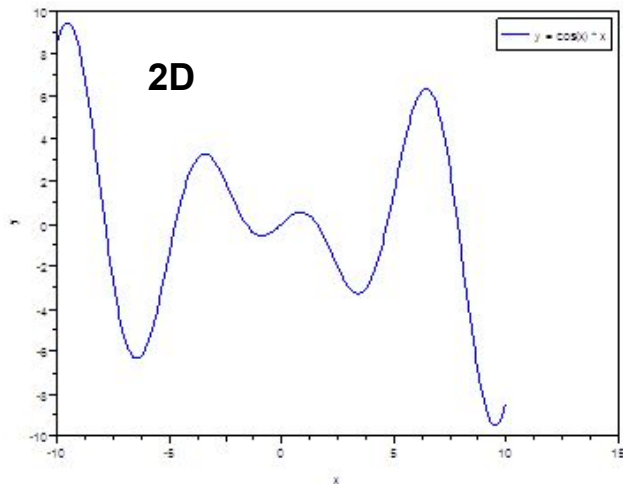
Visualização de dados

Outra possibilidade é analisar a relação entre duas variáveis e representar a variável objetivo pelo tipo de **marcação**



Visualização de dados - TSNE

Seria possível analisar através de um gráfico 2d pontos que estão em um hiperplano de dimensão maior que 2?

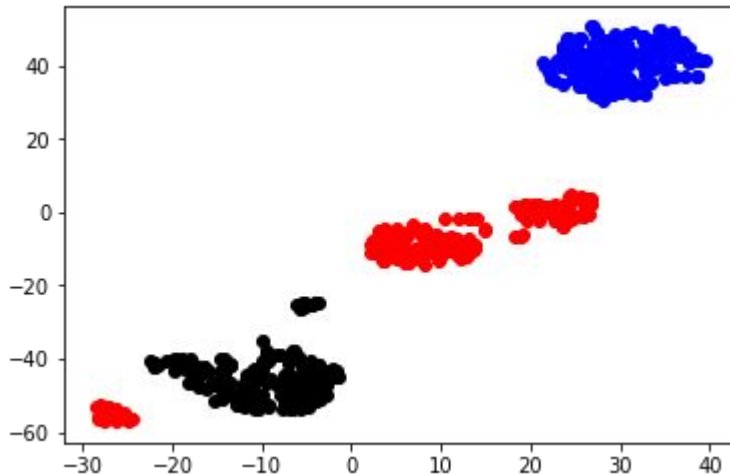


20D



Visualização de dados - TSNE

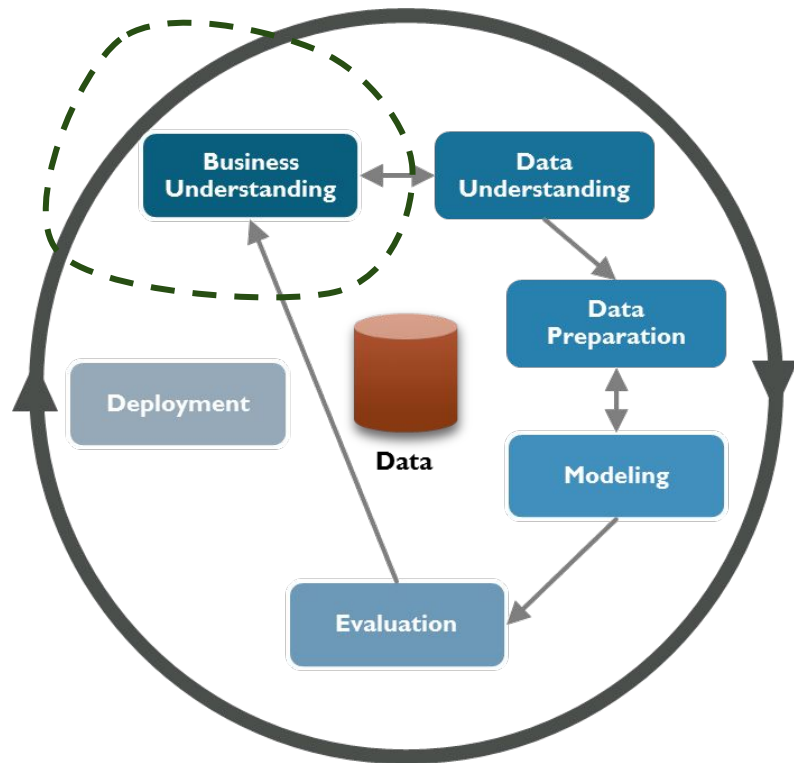
A técnica **TSNE** se propõe a representar um reticulado hiperdimensional por pontos 2d preservando as relações de proximidade. Basicamente com o TSNE podemos ter uma idéia se as features utilizadas são ou não satisfatórias para a classificação



Para o Iris dataset

Como funciona um projeto de Aprendizado de máquina

Neste momento, avaliamos se o desempenho do modelo é adequado ao negócio



CRISP - DM

Por que utilizar Python?

Python é uma linguagem de alto nível **interpretada** e de **propósito geral**.

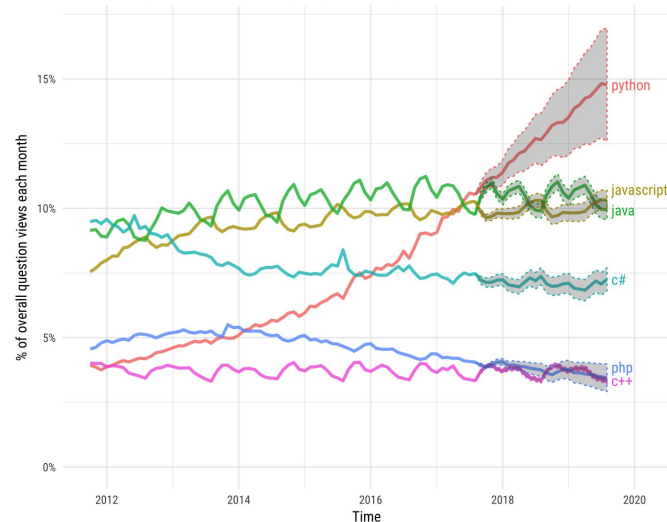
É também **multi-plataforma**, **multi-paradigma**, além de possuir tipagem dinâmica e gerenciamento automático de memória.

É uma das linguagens **mais utilizadas** para aplicações de aprendizado de máquina



Projections of future traffic for major programming languages

Future traffic is predicted with an STL model, along with an 80% prediction interval.



Por que utilizar Python?

Python é uma linguagem de **fácil leitura** e possui muitos pacotes para lidar com **diversos tipos de dados** de forma simples

Imagem



Áudio

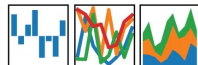


Texto



Dados tabulares

pandas
 $y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$



Por que utilizar Python?

O mesmo é verdade para pacotes de análise de dados

Análise numérica



Visualização de dados



Mineração e leitura de dados



Aprendizado de máquina



Deep Learning



Testando Python no Browser

Ao longo do curso, utilizaremos uma ferramenta simples para testar o Python e suas principais bibliotecas em um ambiente visual



Gerenciando dependências com o Python

O Python possui um gerenciador de dependências bastante simples chamado **pip** os principais comandos são:

pip list

pip install <package>

pip install <package>==<version>

pip uninstall <package>

pip install -r requirements.txt

É possível criar um documento com as dependências da seguinte forma:

```
numpy == 1.14.5
pandas == 0.23.1
scikit-learn == 0.19.1
scipy == 1.1.0
python-dateutil == 2.7.3
tqdm == 4.23.4
pydotplus == 2.0.2
sphinx == 1.7.6
matplotlib == 2.2.2
vertica-python == 0.7.3
s3io == 0.1.1
boto3 == 1.9.11
awscli == 1.16.21
torch == 0.4.0
torchvision == 0.2.1
```

Análise numérica em Python

Os conhecimentos matemáticos mais importantes para a análise de dados são a álgebra linear e a estatística. Em função disso, surgiu a biblioteca **Numpy**

Avalie a primeira parte do curso

https://docs.google.com/forms/d/e/1FAIpQLScjRfErajmoXclnExnMia32RJ9NLDQbtS_DJ25jGHYDSmhQbg/viewform?usp=sf_link