

# Aprendizado de máquina

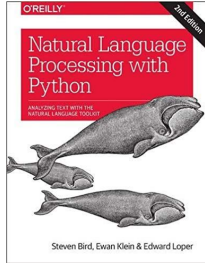
Fundamentos e aplicações em processamento de linguagem natural

Felipe Navarro Balbino Alves

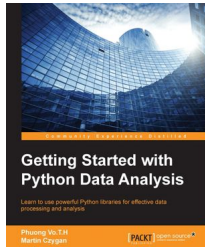
# Github do curso

[https://github.com/fnbalves/curso\\_machine\\_learning/](https://github.com/fnbalves/curso_machine_learning/)

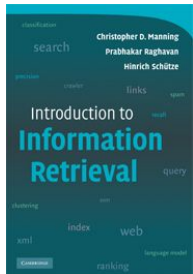
# Bibliografia de interesse



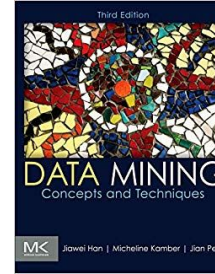
**Natural Language Processing with Python**  
Steven Bird (Disponível online)



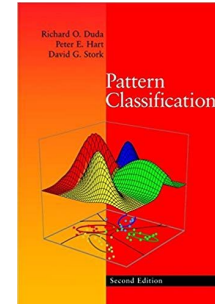
**Getting started with Python Data Analysis**  
Phuong Vo. T.H



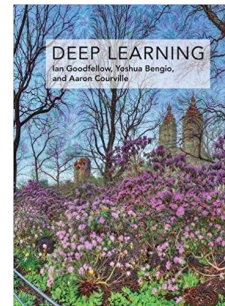
**Introduction to information Retrieval**  
Cristopher D. Manning



**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:

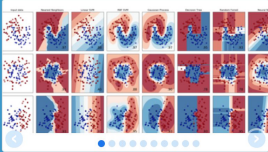
<https://nlp.stanford.edu/IR-book/>

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

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

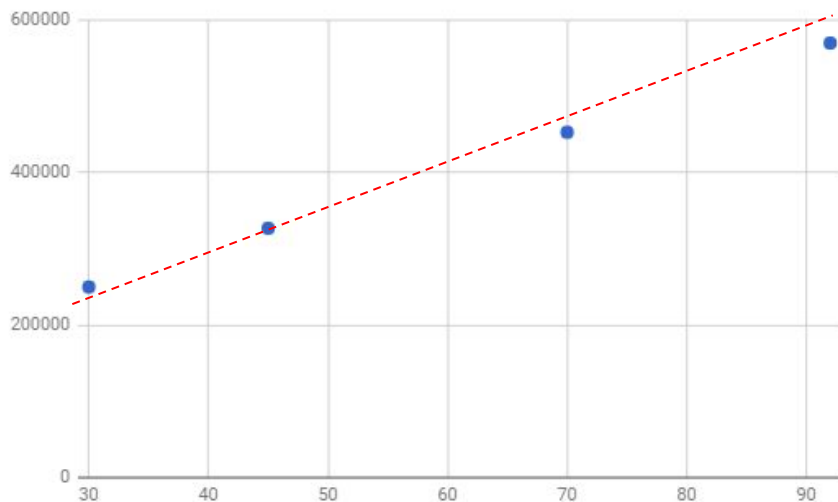
O corretor lhe fornece a seguinte planilha de dados de precificação anteriores

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 <sup>2</sup> )	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 <sup>2</sup> )	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 ▼

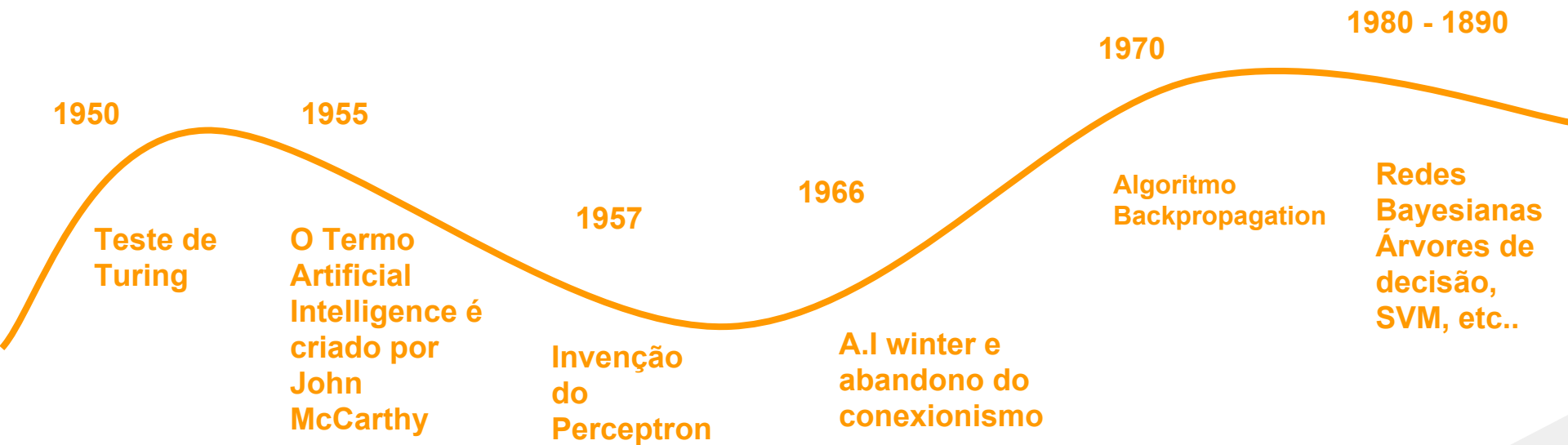
Todas as categorias ▼

Pesquisa na Web ▼

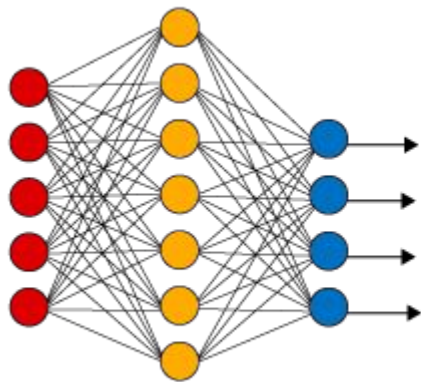
Interesse ao longo do tempo ?



# Evolução do AM

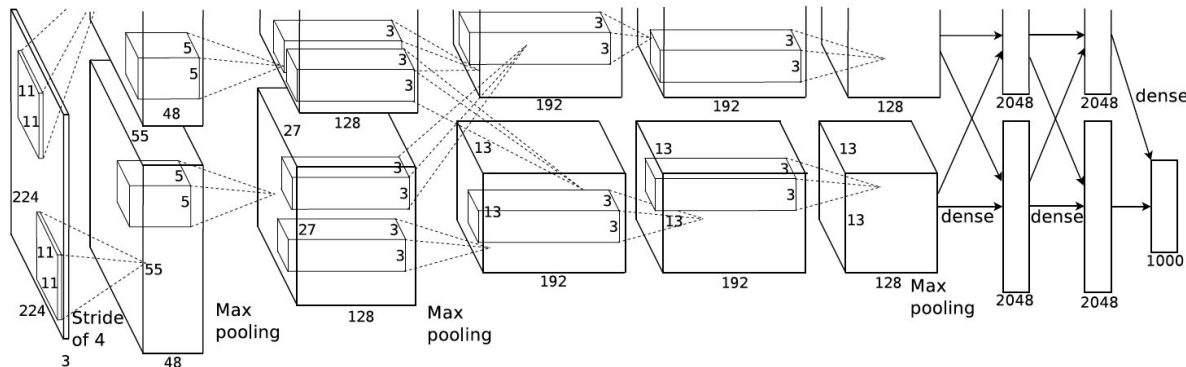


# Surgimento do deep learning

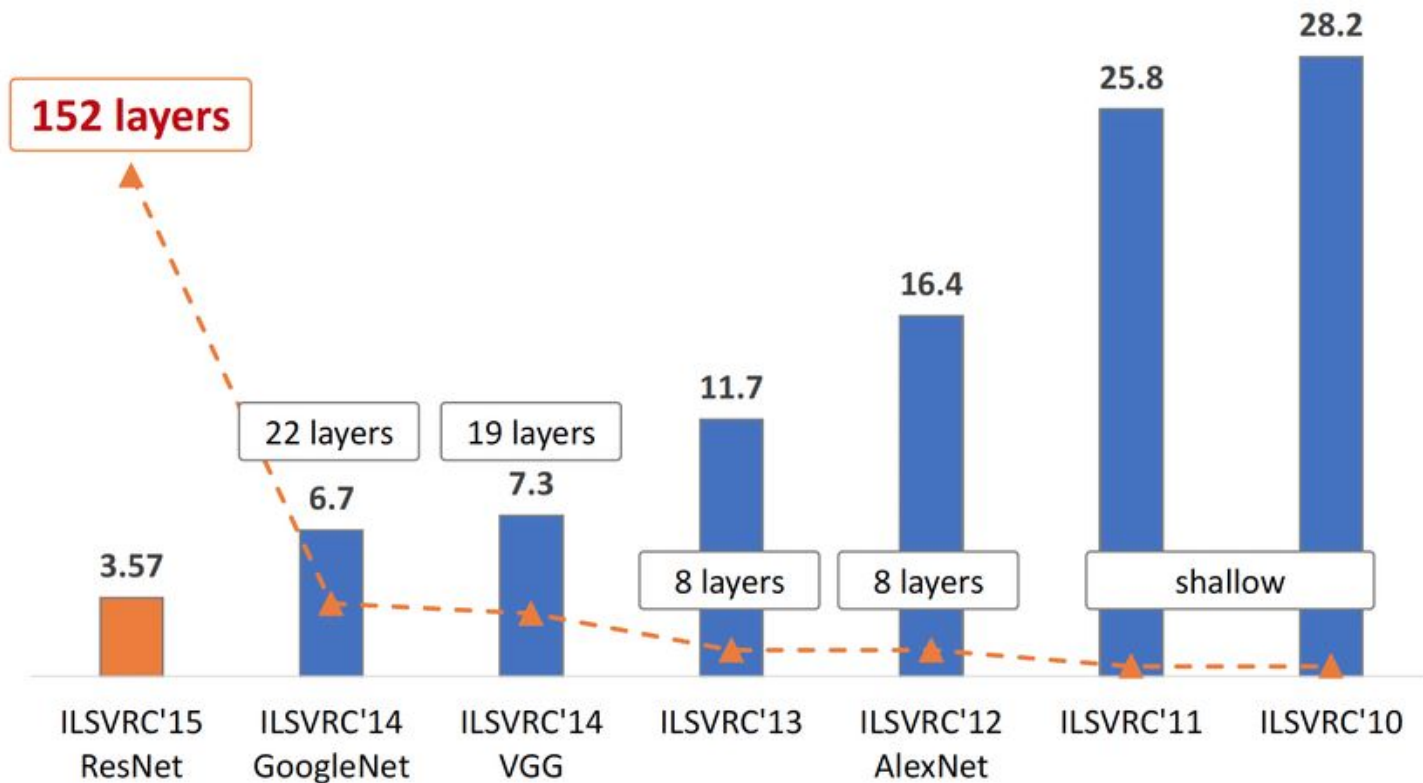


Shallow network

**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

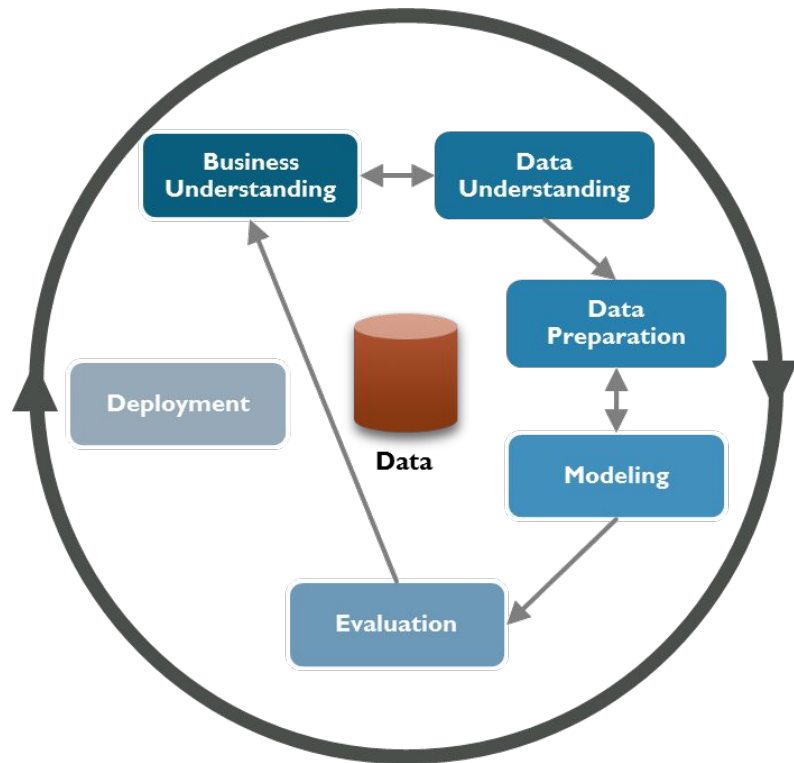


1 Titan Z-Accelerated Server  
3 Titan Zs • 17,280 cores

**2 kWatts**  
**\$12,000**



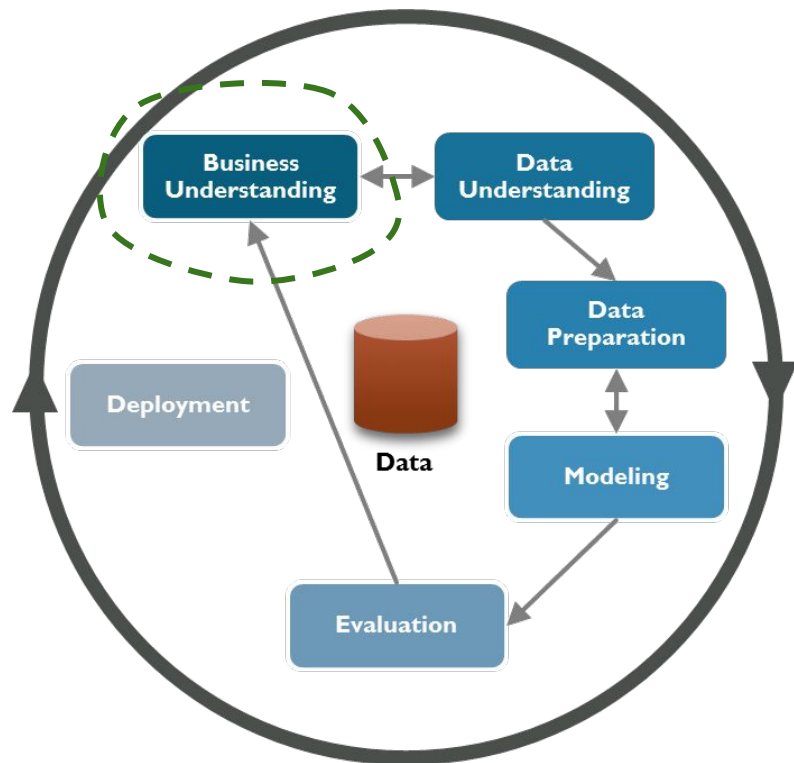
# Como funciona um projeto de Aprendizado de máquina



**CRISP - DM**

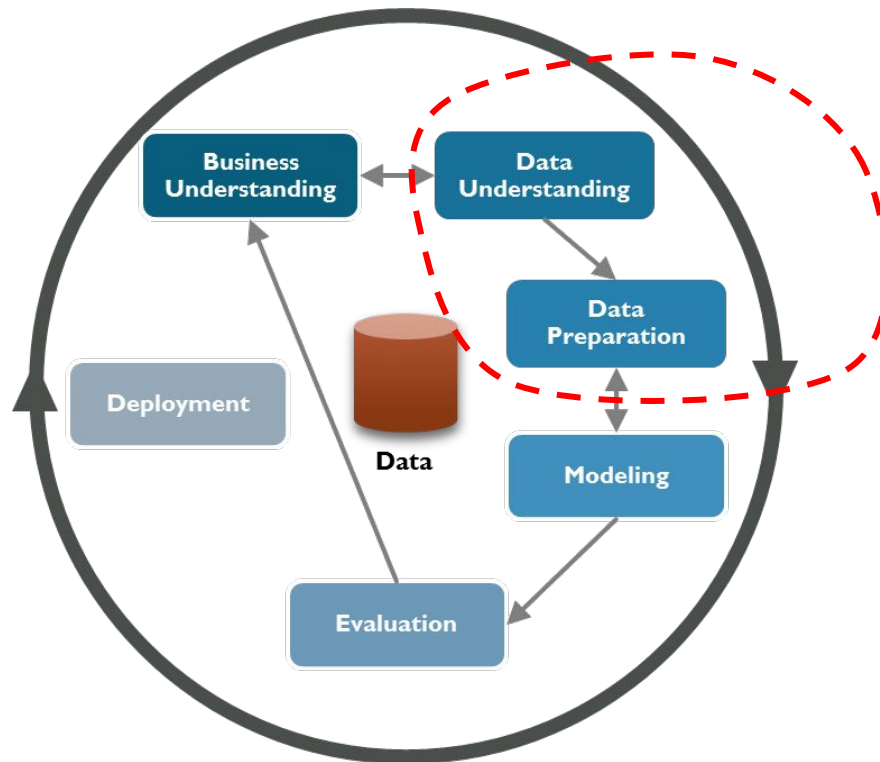
# Como funciona um projeto de Aprendizado de máquina

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 (30m<sup>2</sup>, 49 m<sup>2</sup>, ....)

**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**

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

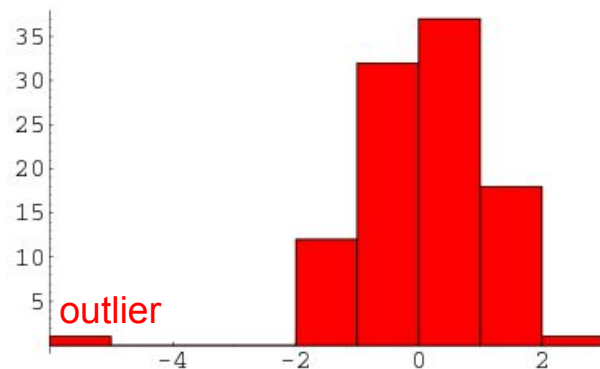
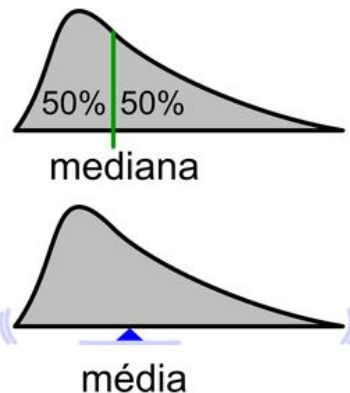
**Variância**

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

**Outliers**

**Mediana**

**Histograma**



# Caracterização estatística dos dados

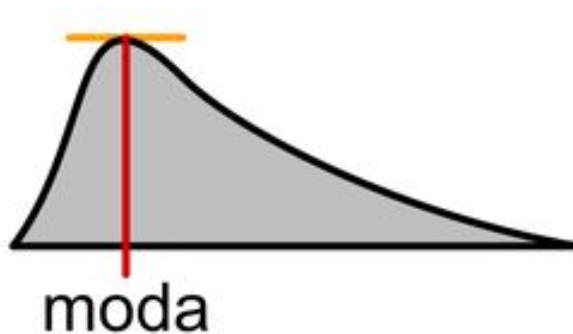
Para **variáveis categóricas**, é importante identificar durante o entendimento dos dados, as seguintes grandezas:

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

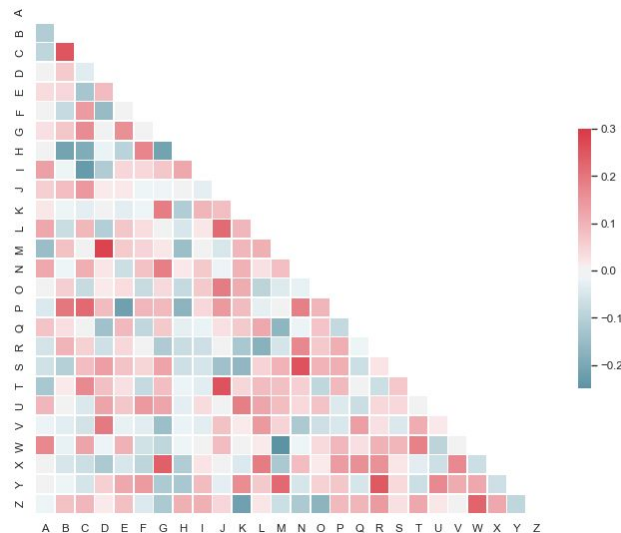
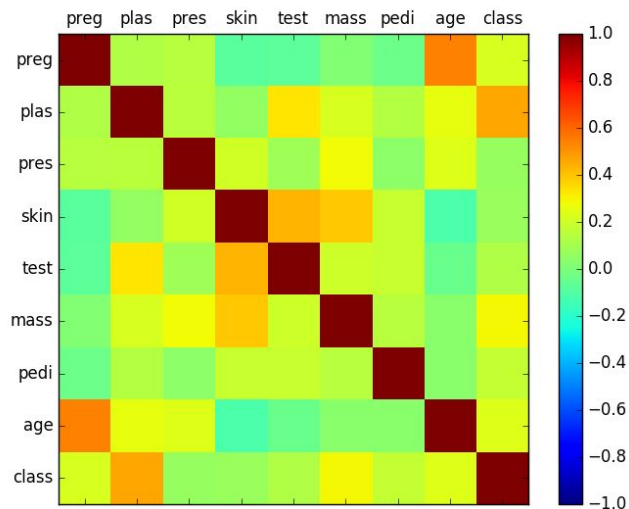
$$\rho = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X) \cdot \text{var}(Y)}}$$

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).

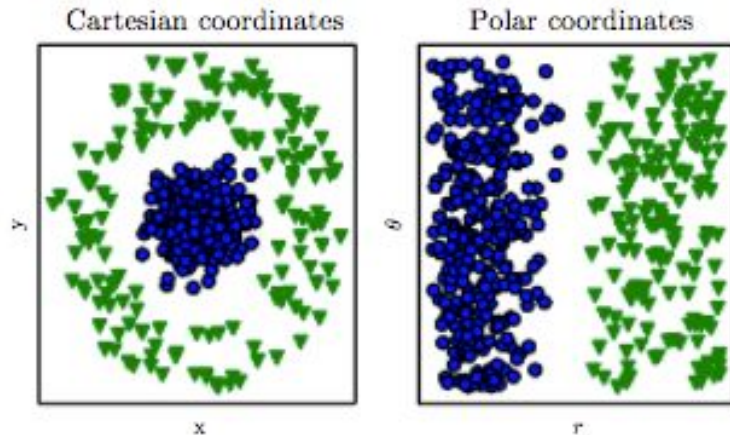
Provavelmente, estas duas variáveis em conjunto não serão mais relevantes do que a variável combinada:

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

# Engenharia de atributos

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

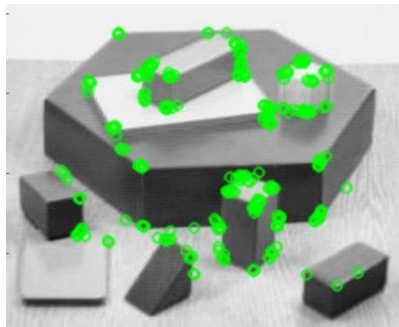
Exemplo 2:



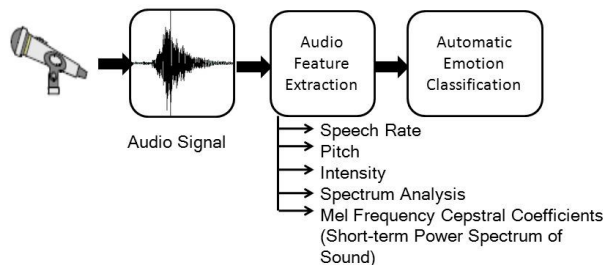
# Engenharia de atributos

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

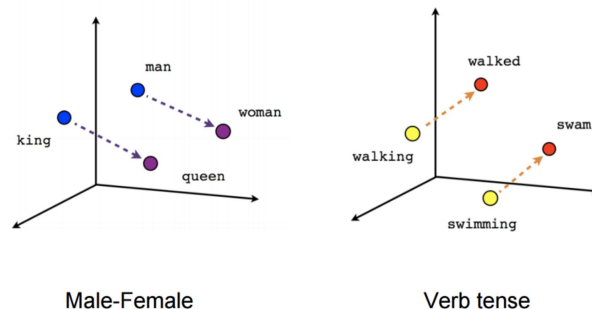
## Processamento de imagens



## 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**

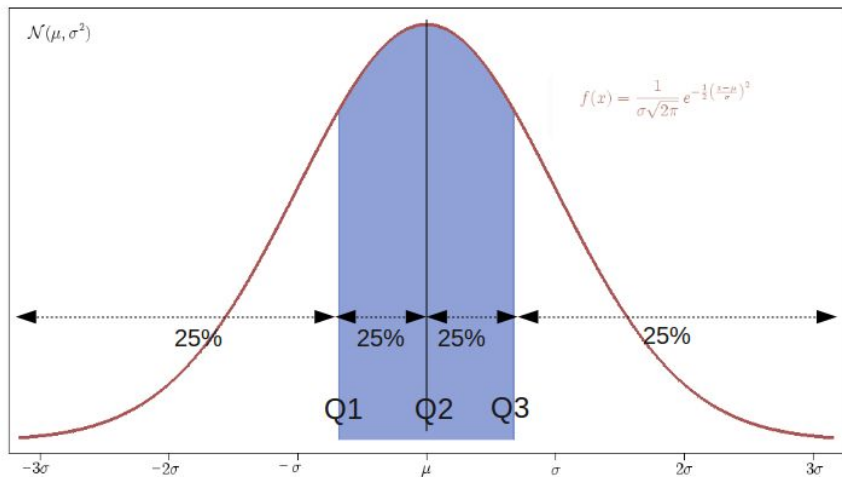
**Moda**

**Quantis**



# Quantis

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

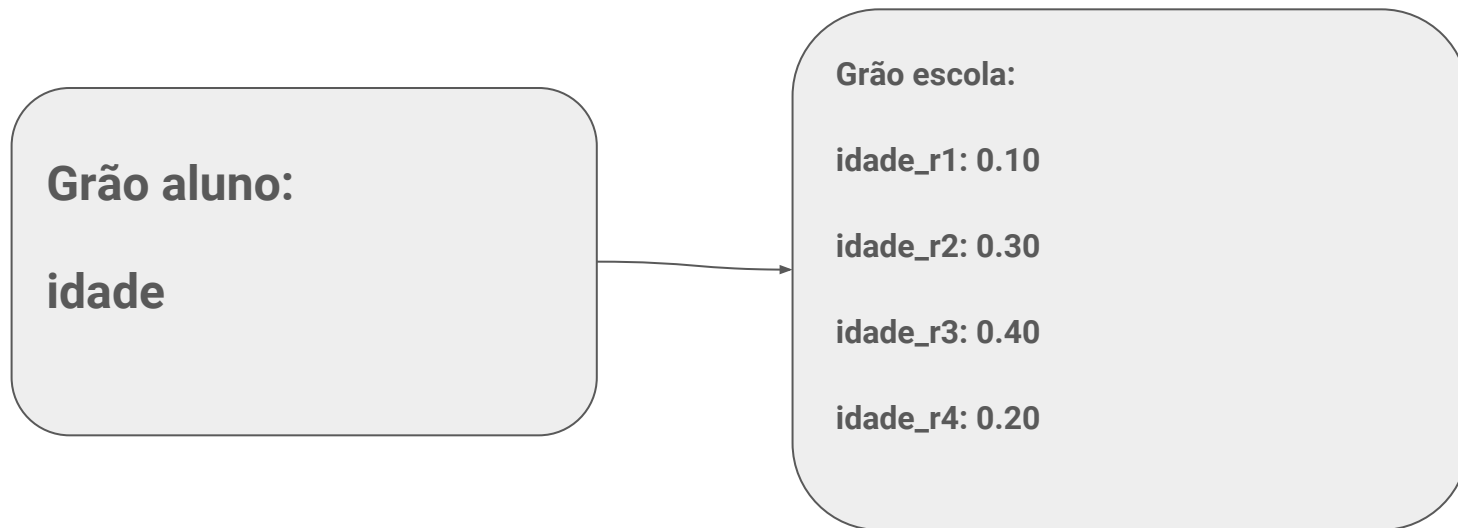


No gráfico maior, podemos colocar a quantidade de elementos que apareceu em cada uma das regiões

# Quantis

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

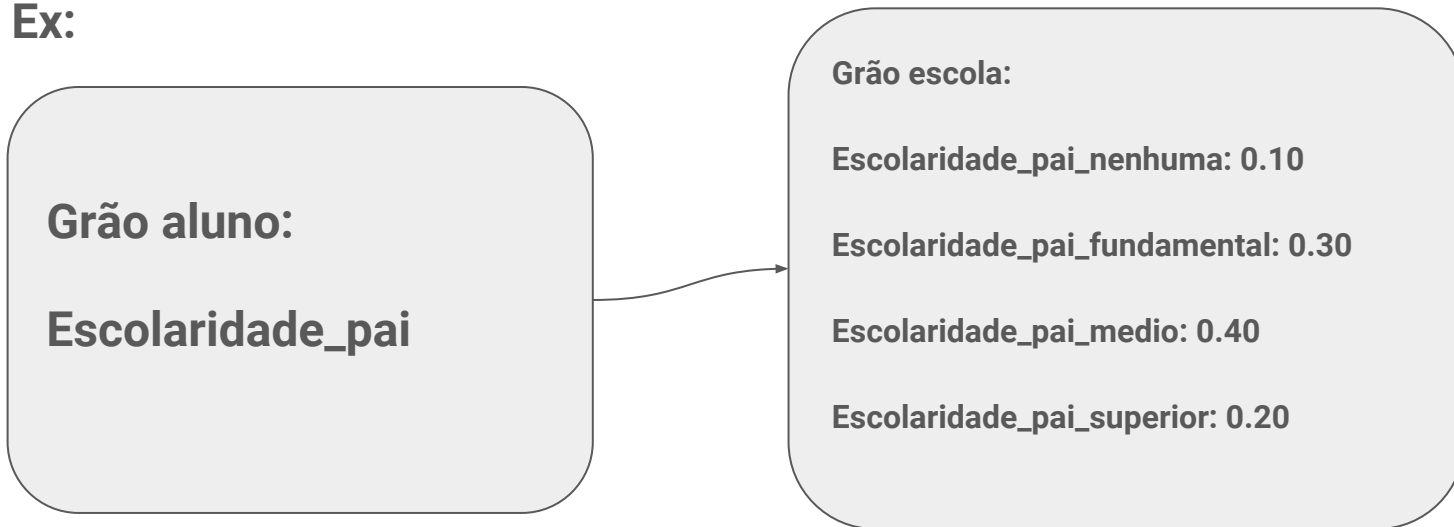
**Ex:**



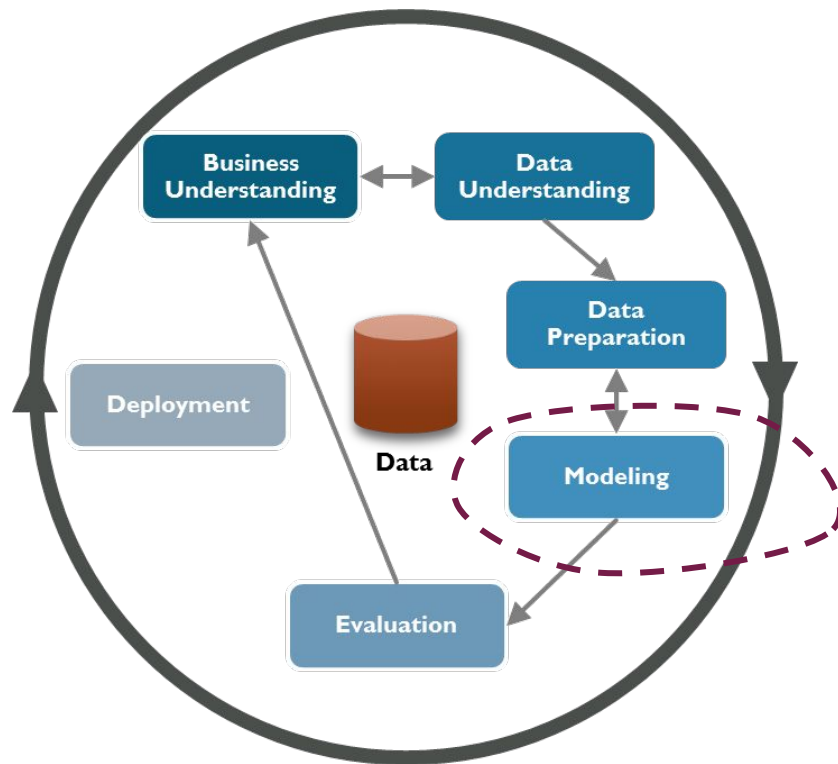
# Mudança de granularidade - variáveis categóricas

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

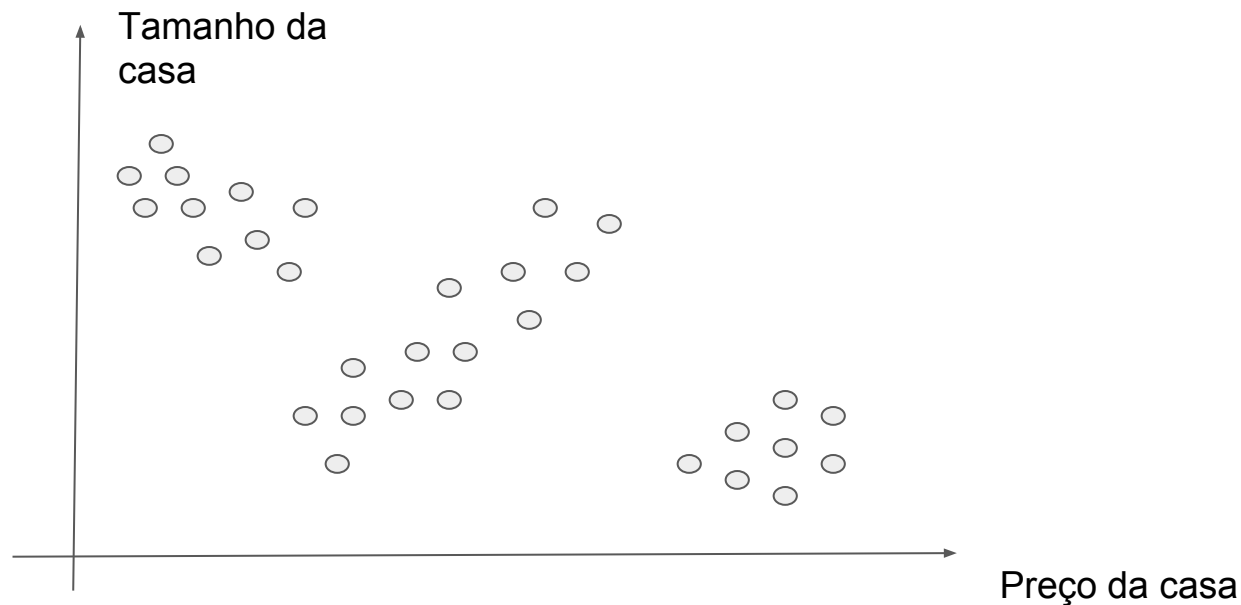


# Aprendizado não supervisionado

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

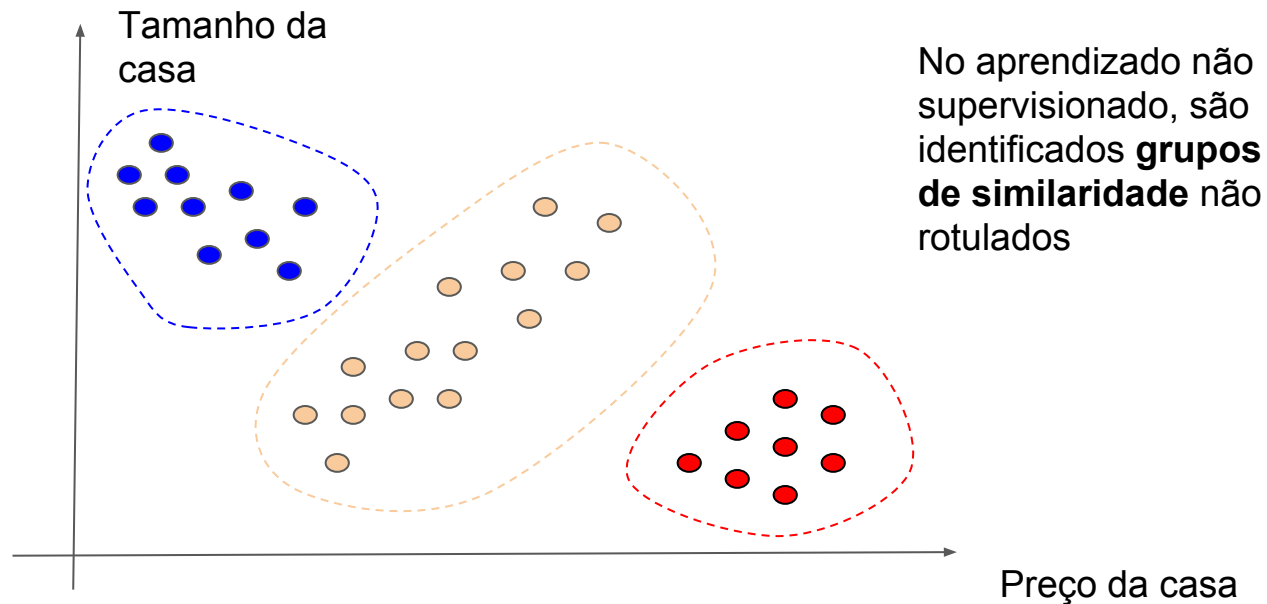
# Aprendizado não supervisionado

## Exemplo



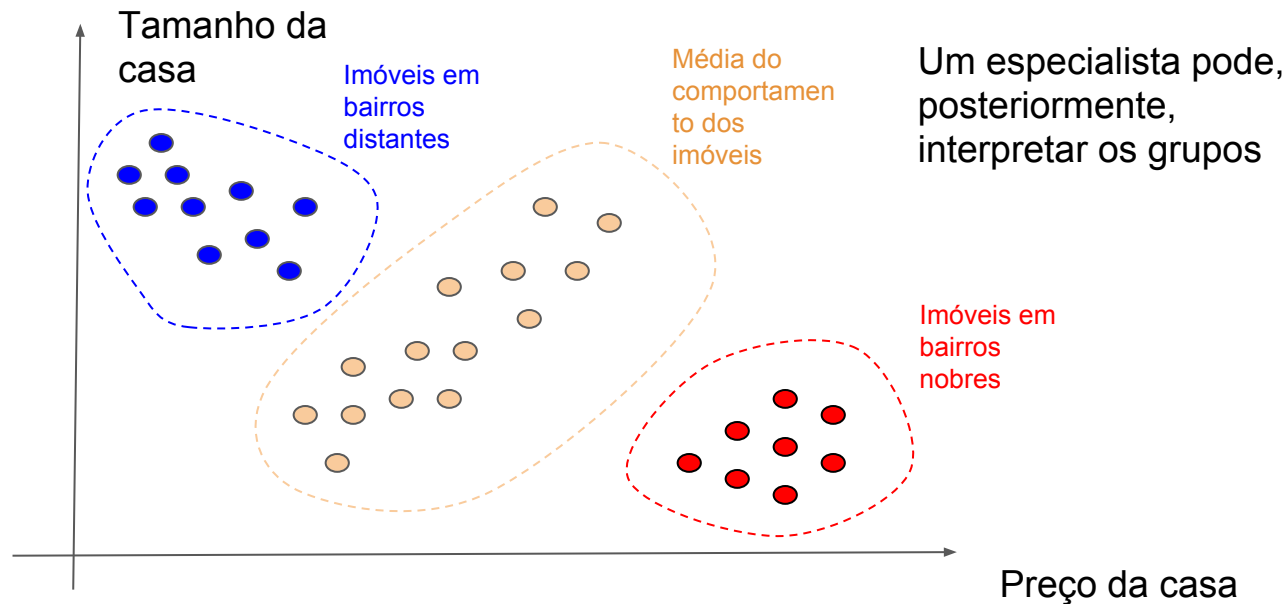
# Formas de aprendizado

## Exemplo



# Formas de aprendizado

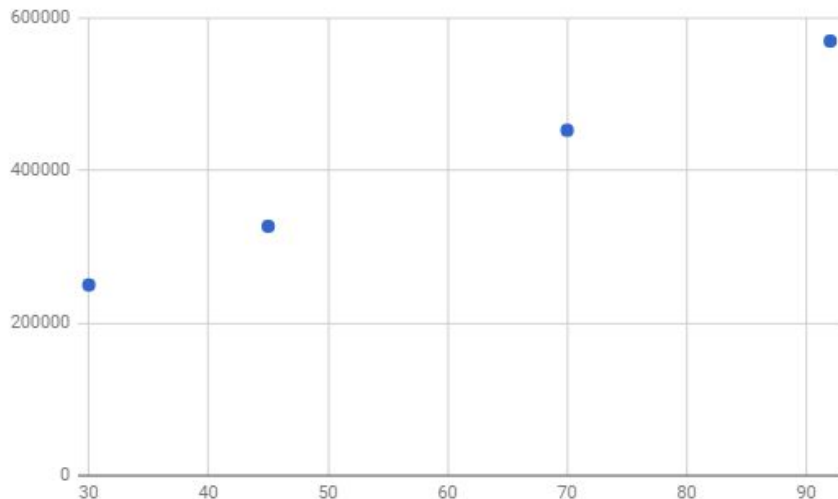
## Exemplo



# Analizando o padrão encontrado

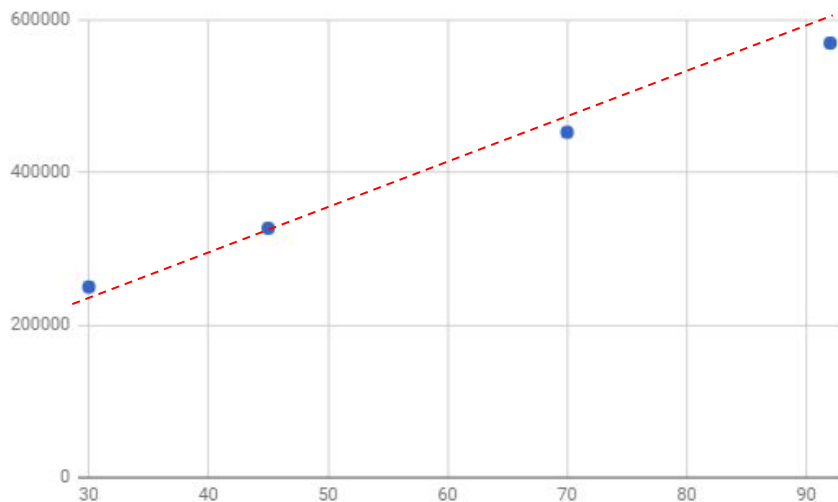
**Exemplo prático:** precificar um imóvel

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



# Avaliando o padrão encontrado

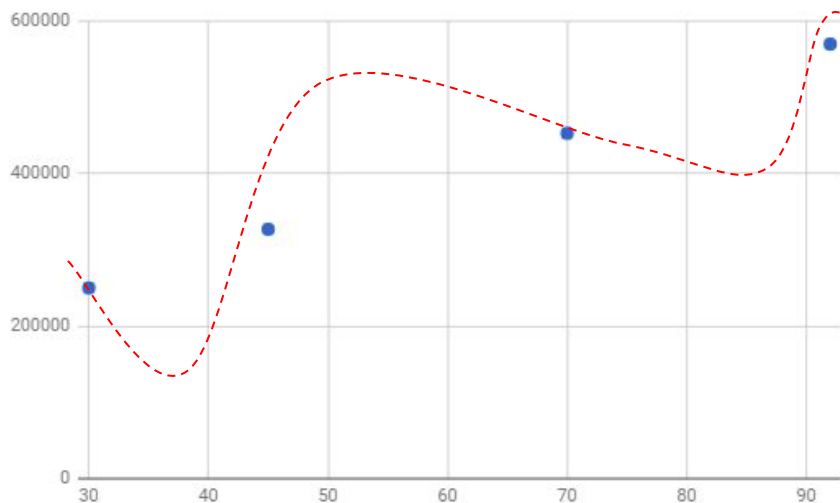
No âmbito da ciência de dados, o padrão também é chamado de **hipótese**



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

# Qual o padrão?

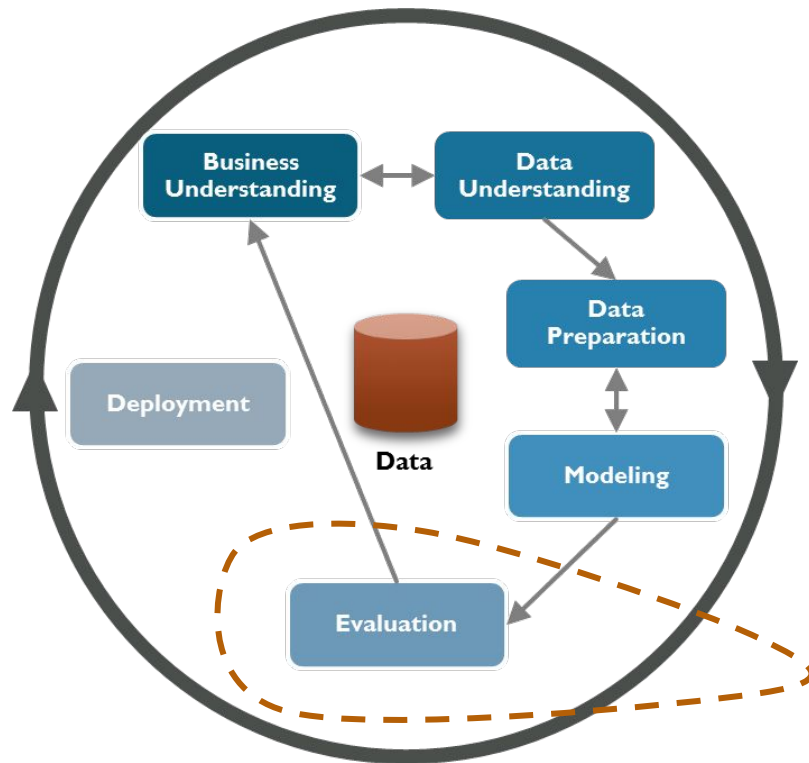
No âmbito da ciência de dados, o padrão também é chamado de **hipótese**



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

Será que ela é melhor?

# Como funciona um projeto de Aprendizado de máquina



**CRISP - DM**

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

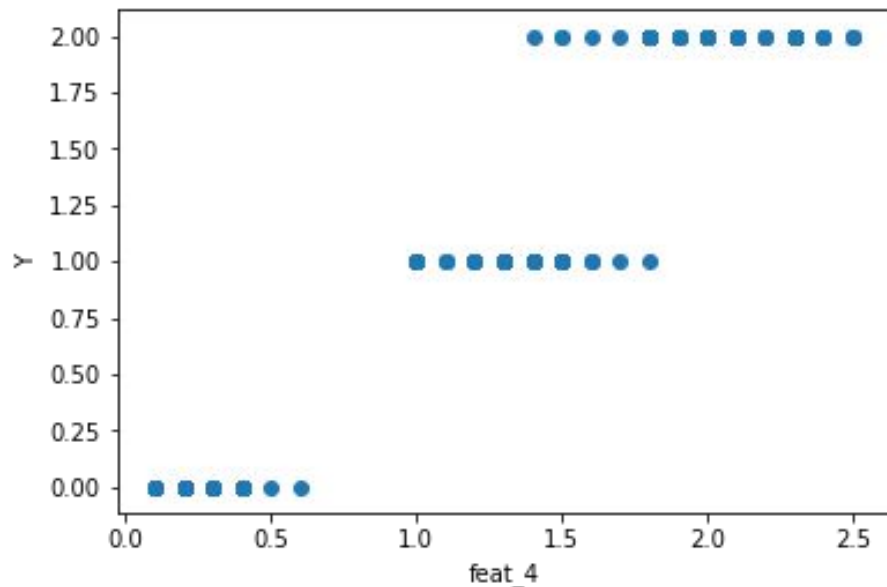


# Visualização de dados

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**

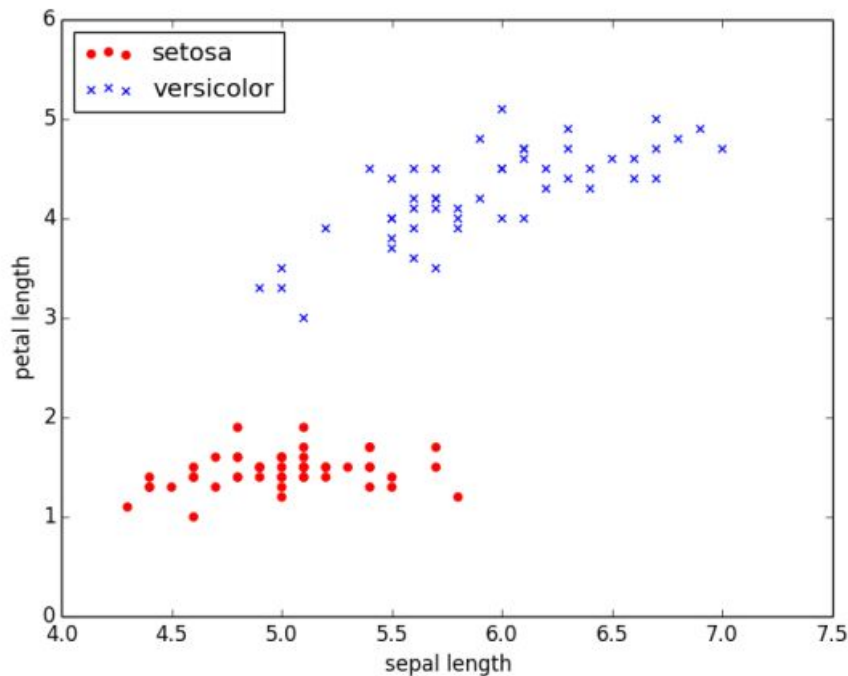
# Visualização de dados

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



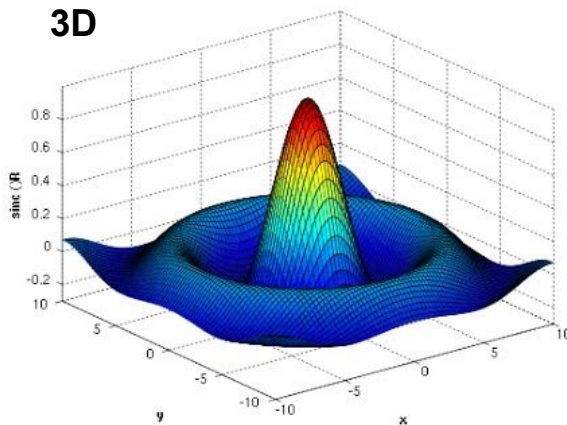
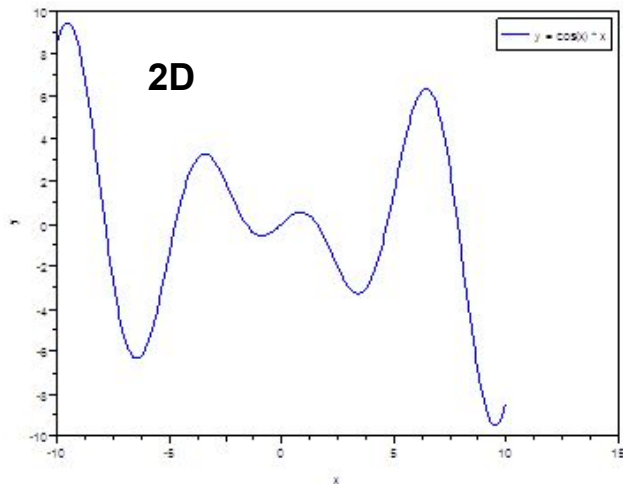
# Visualização de dados

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?

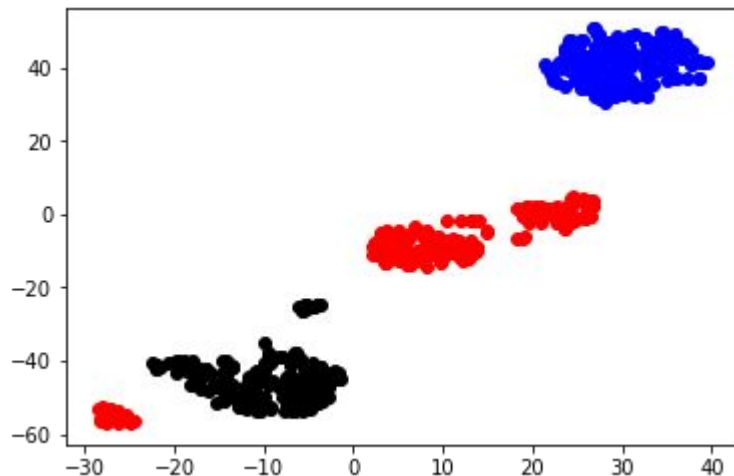


**20D**



# Visualização de dados - TSNE

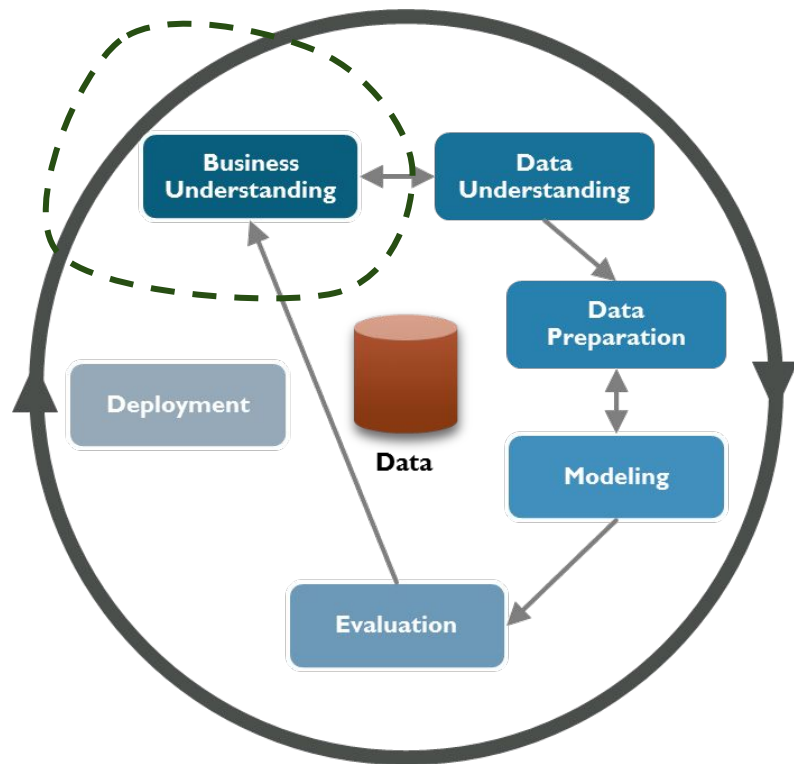
A técnica **TSNE** se propõe a representar um reticulado hiperdimensional por pontos 2d preservando as relações de proximidade. Basicamente com o TSNE podemos ter uma idéia se as features utilizadas são ou não satisfatórias para a classificação



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**.

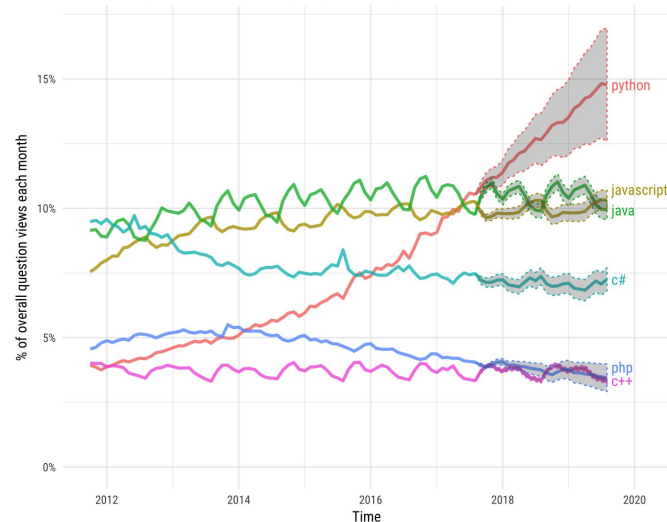
É também **multi-plataforma**, **multi-paradigma**, além de possuir tipagem dinâmica e gerenciamento automático de memória.

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

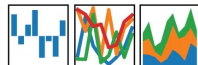


Texto



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**

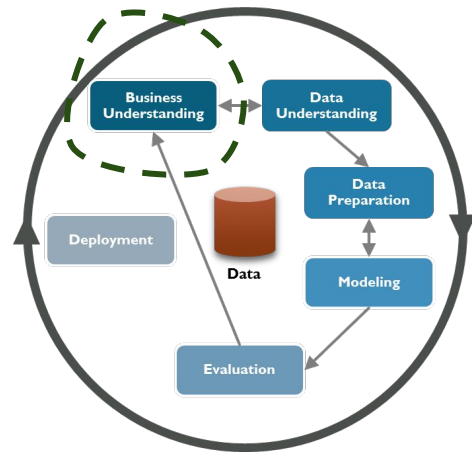
# Exemplo de entendimento dos dados

Análise da qualidade das instituições de ensino superior utilizando microdados do ENADE e do Censo da Educação Superior

# Caracterização do problema

Ao longo dos últimos 20 anos, o número de instituições de educação superior no Brasil mais que dobrou. No entanto:

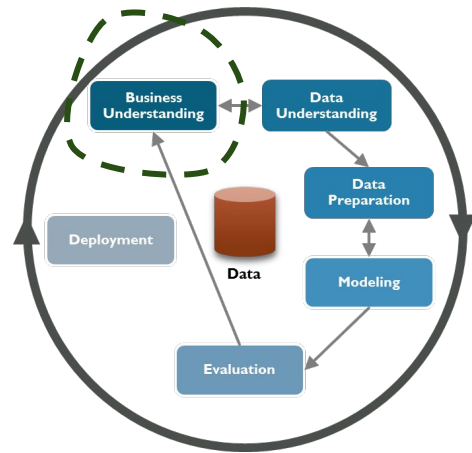
1. O Brasil ainda é o 60º colocado em educação em um ranking de 76 países criado pela OCDE (Organização para a Cooperação e Desenvolvimento Econômico);
2. O Brasil não possui nenhuma universidade entre as 100 melhores do mundo.



# Objetivo

Definir os **fatores que mais influenciam** para a qualidade de um **curso de graduação**.

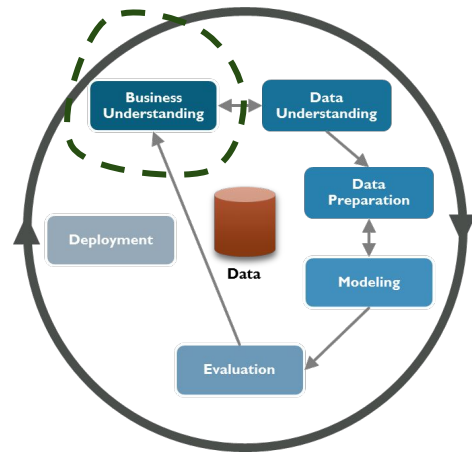
Esses fatores podem influenciar na criação de **diretrizes mais eficientes** para tratar o déficit da educação de nível superior brasileira.



# Entendimento do negócio

Definir os **fatores** que mais influenciam para a qualidade de um **curso de graduação**.

Esses fatores podem influenciar na criação de **diretrizes mais eficientes** para tratar o déficit da educação de nível superior brasileira.





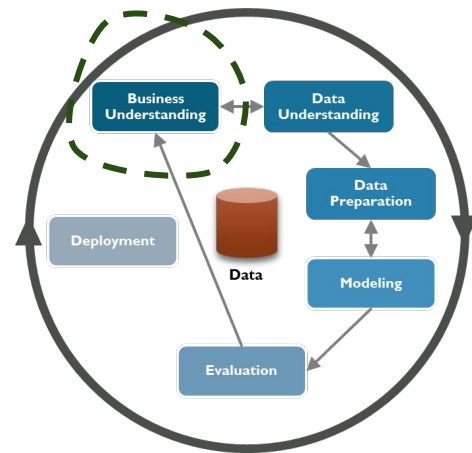
# Entendimento do negócio

## Como as instituições são avaliadas hoje em dia

Atualmente, o INEP (Instituto Nacional de Estudos e Pesquisas Educacionais) utiliza as seguintes métricas para avaliar os cursos de graduação no Brasil.

1. Notas dos alunos no **ENADE**;
2. Características do corpo docente;
3. Instalações físicas;
4. Organização didático-pedagógica;

Essas informações são integradas no chamado índice **CPC** (Conceito preliminar de curso)

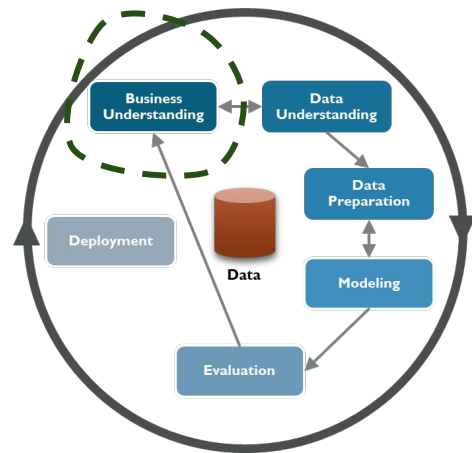


# Entendimento do negócio

## Definição da métrica de desempenho

Queremos verificar também a influência das características do corpo docente e da organização didático-pedagógica na qualidade dos cursos. Dessa forma, a utilização do CPC como métrica de desempenho carrega **informação à posteriori**.

Assim, podemos utilizar apenas a nota do **ENADE**, uma vez que ele **já é utilizado** para avaliar as escolas implicitamente no CPC.



# Entendimento do negócio

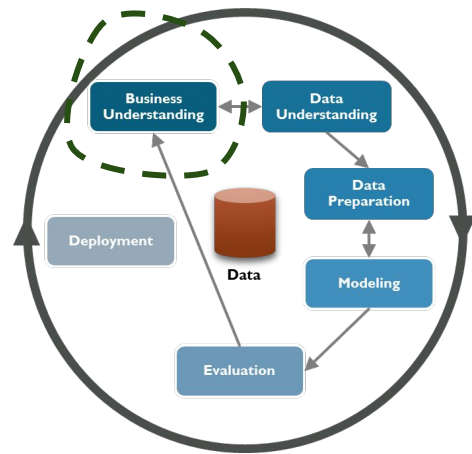
## ENADE

Prova destinada à avaliação de instituições de ensino superior, obrigatória para os alunos selecionados e **condição indispensável** para a emissão de um histórico escolar.

O ENADE possui um **Ciclo de Avaliação**, onde a cada ano, apenas algumas áreas do conhecimento se submetem à prova.

Últimos dados disponíveis: ENADE 2014

Avaliaremos apenas as áreas do conhecimento que **prestaram a prova em 2014**.



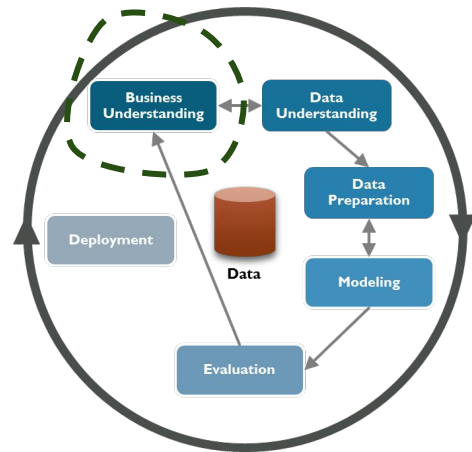
# Entendimento do negócio

## Definição do grão de análise

Assim como o INEP, queremos avaliar as instituições respeitando as **diferenças entre os cursos**.

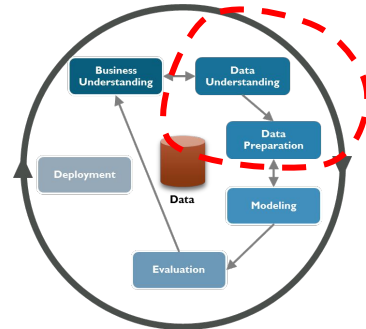
**GRÃO DE ANÁLISE (VERSÃO 1):** O par Instituição x curso, exemplo:

UFPE - Engenharia Eletrônica; UFBA - Ciência da Computação



# Entendimento dos dados

Base de dados disponível



## Censo Escolar

**Cerca de 5GB de dados**

Dados de várias granularidades

- Instituição: 2.368
- Curso: 33.274
- Aluno 10.793.933
- Professor: 396.596

## ENADE

**481.721 registros**

Dados na granularidade aluno

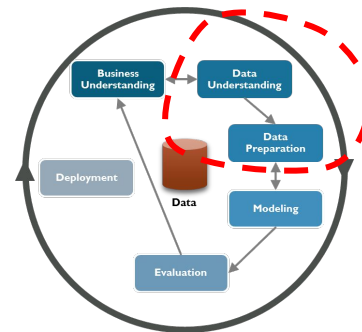
- Existe um índice comum entre as duas bases de dados para identificar a instituição de ensino (IES);
- Não existe um índice comum para a identificação do curso;
- Nas duas bases de dados, existem campos que identificam a **área do curso**;

**GRÃO DE ANÁLISE (Versão 2):** Instituição x área do conhecimento:

Ex : UFPE - Humanas / UFPE - Engenharia / USP - Educação, etc....

# Entendimento dos dados

Área dos cursos - Base do Censo Escolar



São utilizados códigos OCDE (Organização para Cooperação e Desenvolvimento Econômico). O primeiro dígito define a área geral do conhecimento:

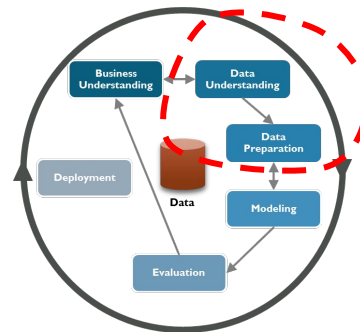
1	Educação
2	Humanidades e Artes
3	Ciências Sociais, Negócios e Direito
4	Ciências, Matemáticas e Computação
5	Engenharia, Produção e Construção
6	Agricultura e Veterinária
7	Saúde e bem estar social
8	Serviços

# Entendimento dos dados

## Área do curso - Base do ENADE

21 = AROUITETURA E URBANISMO  
 72 = TECNOLOGIA EM ANÁLISE E DESENVOLVIMENTO DE SISTEMAS  
 73 = TECNOLOGIA EM AUTOMAÇÃO INDUSTRIAL  
 76 = TECNOLOGIA EM GESTÃO DA PRODUÇÃO INDUSTRIAL  
 79 = TECNOLOGIA EM REDES DE COMPUTADORES  
 701 = MATEMÁTICA (BACHARELADO)  
 702 = MATEMÁTICA (LICENCIATURA)  
 903 = LETRAS-PORTUGUÊS (BACHARELADO)  
 904 = LETRAS-PORTUGUÊS (LICENCIATURA)  
 905 = LETRAS-PORTUGUÊS E INGLÊS (LICENCIATURA)  
 906 = LETRAS-PORTUGUÊS E ESPANHOL (LICENCIATURA)  
 1401 = FÍSICA (BACHARELADO)  
 1402 = FÍSICA (LICENCIATURA)  
 1501 = QUÍMICA (BACHARELADO)  
 1502 = QUÍMICA (LICENCIATURA)  
 1601 = CIÊNCIAS BIOLÓGICAS (BACHARELADO)  
 1602 = CIÊNCIAS BIOLÓGICAS (LICENCIATURA)  
 2001 = PEDAGOGIA (LICENCIATURA)  
 2401 = HISTÓRIA (BACHARELADO)  
 2402 = HISTÓRIA (LICENCIATURA)  
 2501 = ARTES VISUAIS (LICENCIATURA)  
 3001 = GEOGRAFIA (BACHARELADO)  
 3002 = GEOGRAFIA (LICENCIATURA)  
 3201 = FILOSOFIA (BACHARELADO)  
 3202 = FILOSOFIA (LICENCIATURA)  
 3502 = EDUCAÇÃO FÍSICA (LICENCIATURA)  
 4004 = CIÊNCIA DA COMPUTAÇÃO (BACHARELADO)  
 4005 = CIÊNCIA DA COMPUTAÇÃO (LICENCIATURA)  
 4006 = SISTEMAS DE INFORMAÇÃO  
 4301 = MÚSICA (LICENCIATURA)  
 5401 = CIÊNCIAS SOCIAIS (BACHARELADO)  
 5402 = CIÊNCIAS SOCIAIS (LICENCIATURA)  
 5710 = ENGENHARIA CIVIL  
 5806 = ENGENHARIA ELÉTRICA  
 5809 = ENGENHARIA DE COMPUTAÇÃO  
 5814 = ENGENHARIA DE CONTROLE E AUTOMAÇÃO  
 5902 = ENGENHARIA MECÂNICA  
 6008 = ENGENHARIA QUÍMICA  
 6009 = ENGENHARIA DE ALIMENTOS  
 6208 = ENGENHARIA DE PRODUÇÃO  
 6306 = ENGENHARIA  
 6307 = ENGENHARIA AMBIENTAL  
 6405 = ENGENHARIA FLORESTAL

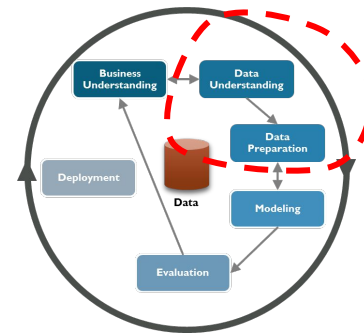
Realizamos a conversão dos códigos utilizados no ENADE para os códigos OCDE através de um dicionário.



# Entendimento dos dados

## Base do Censo Escolar

Esta base possui dados em vários grãos



**DM\_ALUNO:** Grão aluno,  
Identificação dos cursos/ área/ sexo / raça / idade / necessidades especiais / participação em pesquisas ou projetos assistencialistas / etc...;

**DM\_CURSO:** Grão curso,  
Identificação dos cursos/ área/ carga horária/ recursos didáticos e acessibilidade/ vagas e métodos de ingresso/ número de concluintes e ingressantes;

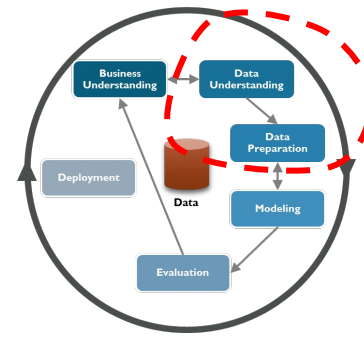
**DM\_DOCENTE:** Grão docente (está separado por IES, não por curso)  
Escolaridade / necessidades especiais;

**DM\_IES:** Grão instituição de ensino;  
Código da Instituição / Quantidade de Técnicos de todos os níveis / Valores de Receitas (Transferências) em R\$



# Entendimento dos dados

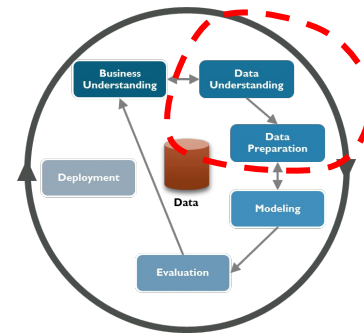
Base ENADE



Grão aluno. Utilizamos apenas a informação da nota média, pois os demais dados já estão presentes em DM\_ALUNO com melhor qualidade

# Pré processamento dos dados

Missing data

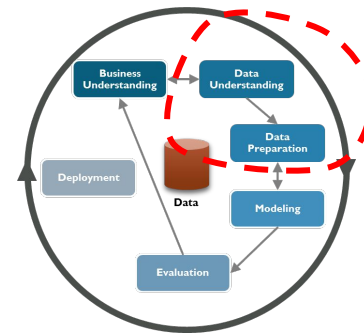


**Missing data para valores categóricos:** Criação de uma nova categoria para missing;

**Missing para valores numéricos:** Substituição pela média dos demais (variância artificial)

# Pré processamento dos dados

## Aglutinação dos dados



**Variáveis categóricas:** Criação de variáveis dummy e aglutinação pela média, assim, no grão final, temos a proporção de elementos em cada uma das classes;

**Valores monetários:** Aglutinação pela média;

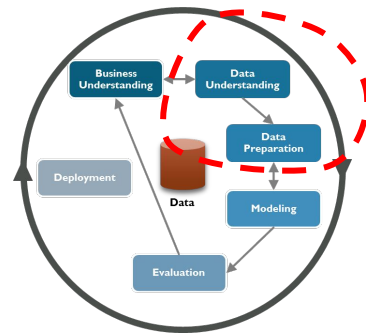
**Variáveis numéricas que representam contagens ou valores contínuos** (ex: Número de vagas EAD ou idade do aluno): Aglutinação pela média;

**Datas** (ex: **Data de abertura do curso**): Consideramos apenas o ano e tomamos a média;

Idade média = Média(ano atual - ano de abertura) = ano atual - Média (ano de abertura)

# Pré processamento dos dados

## Colunas removidas



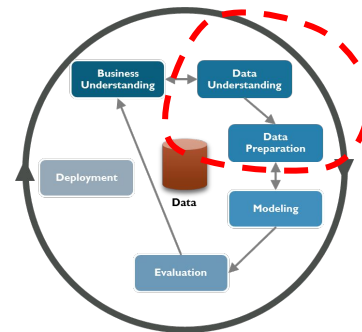
A fim de garantir que o classificador construído não aprenda aspectos regionais da educação, removemos as colunas relativas à localização das IES, ex: UF, Estado, etc...

Para alunos estrangeiros, consideramos apenas a booleana: é brasileiro ou não, ignoramos o país de origem por possuir muitas categorias

# Pré processamento dos dados

Mudança de grão e filtragem de dados

DM\_ALUNO



10793935 (total)

6809245 (aluno cursando)

6786979  
(co\_ocde\_area\_geral  
presente)

**perda total: 37.12 %**

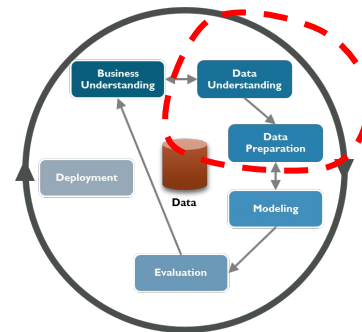
agrupamento  
IES / Área

7222  
registros

# Pré processamento dos dados

Mudança de grão e filtragem de dados

DM\_CURSO



33273 (total)

31069 (curso ativo e com ano de início)

30868(co\_ocde\_area\_geral presente)

**perda total: 7.22 %**

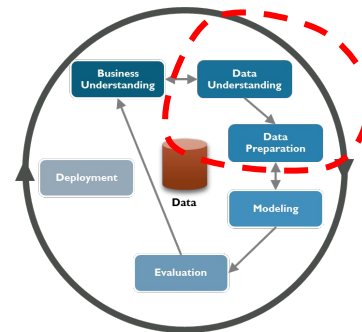
agrupamento  
IES / Área

7215  
registros

# Pré processamento dos dados

Mudança de grão e filtragem de dados

DM\_DOCENTE



396595 (total)

383386 (docente em atividade)

**perda total: 3.33 %**

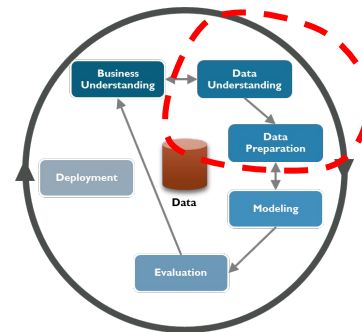
agrupamento  
IES

2368  
registros

# Pré processamento dos dados

Mudança de grão e filtragem de dados

DM\_ENADE



481720 (total)

395557 (presente)

395453 (nota geral  
presente)

**perda total: 17.90 %**

agrupamento  
IES / Área

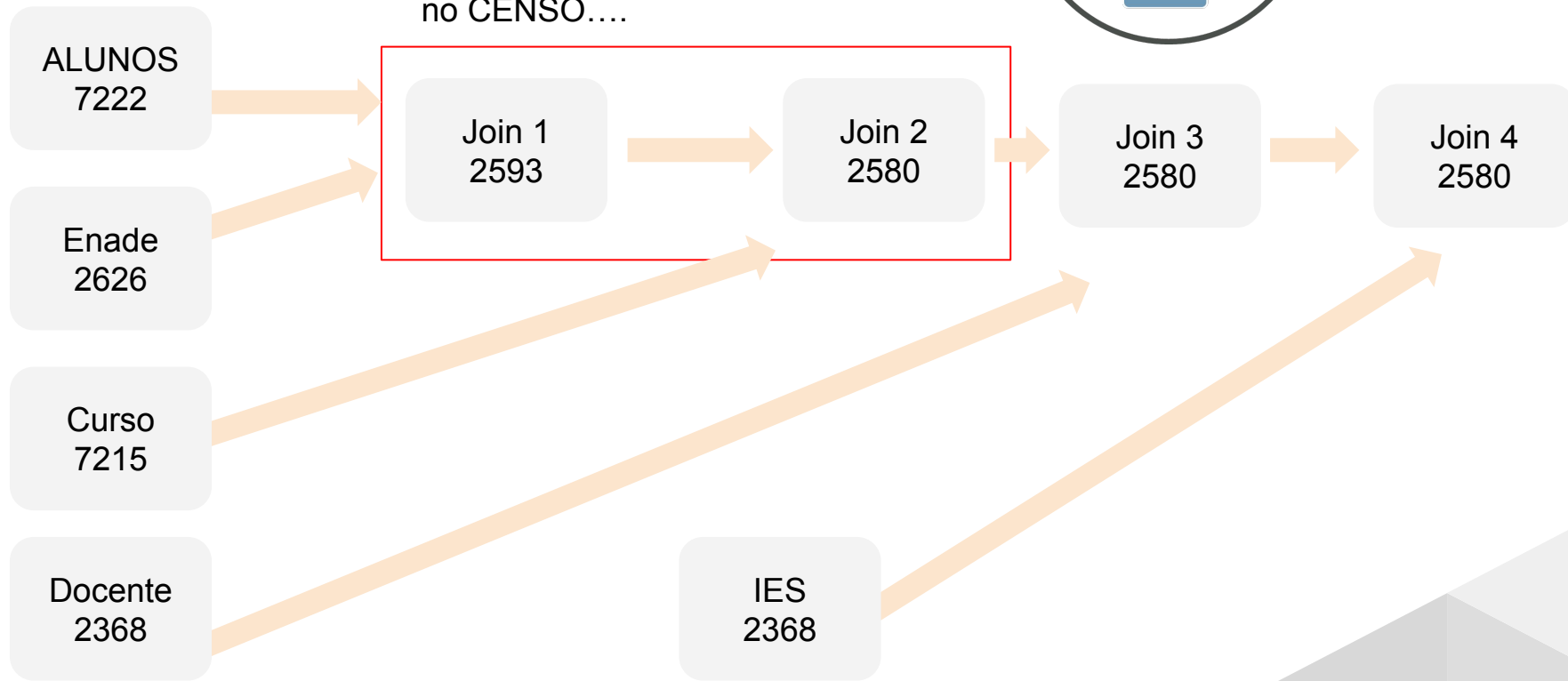
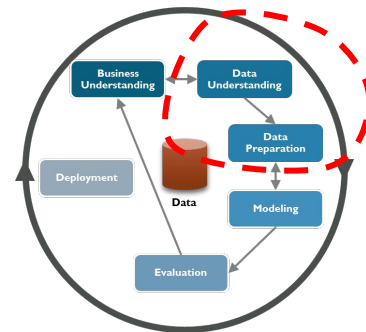
2626  
registros



# Pré processamento dos dados

## Junção de dados

IES presentes no Enade mas não no CENSO....



## Avalie a primeira parte do curso

[https://docs.google.com/forms/d/e/1FAIpQLScjRfErajmoXclnExnMia32RJ9NLDQbtS\\_DJ25jGHYDSmhQbg/viewform?usp=sf\\_link](https://docs.google.com/forms/d/e/1FAIpQLScjRfErajmoXclnExnMia32RJ9NLDQbtS_DJ25jGHYDSmhQbg/viewform?usp=sf_link)