

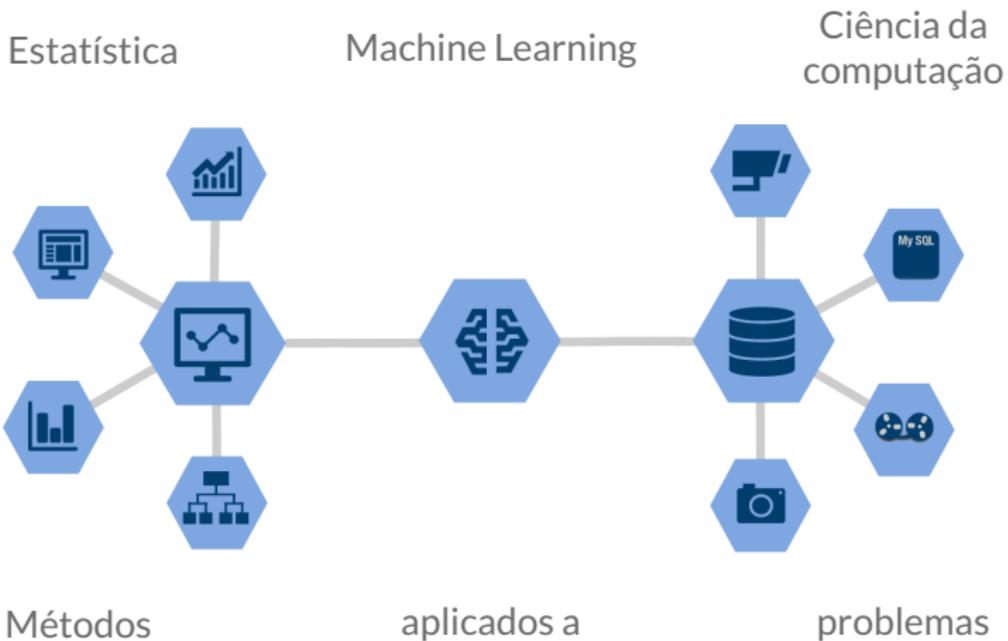
# Introdução

Prof. Eduardo Vargas Ferreira

Curso de Especialização em  
Data Science & Big Data  
Universidade Federal do Paraná

10 de agosto de 2018

# O que é Machine Learning?



# Predição versus Inferência

## ► Data Modeling Culture

- ▶ Domina a comunidade estatística;
- ▶ O principal objetivo está na interpretação dos parâmetros;
- ▶ Testar suposições é fundamental.

## ► Algorithmic Modeling Culture

- ▶ Domina a comunidade de Machine Learning;
- ▶ O modelo é utilizado para criar bons algoritmos preditivos;
- ▶ Interpretamos os resultados, mas esse - em geral - não é o foco.

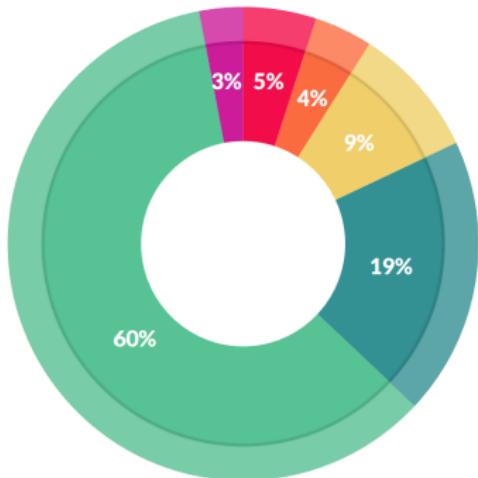
L. Breiman. Statistical modeling: The two cultures. *Statistical Science*, 16(3):199-231, 2001



# Algoritmos de Machine Learning



# Onde desprendemos mais tempo



What data scientists spend the most time doing

- *Building training sets:* 3%
- *Cleaning and organizing data:* 60%
- *Collecting data sets:* 19%
- *Mining data for patterns:* 9%
- *Refining algorithms:* 4%
- *Other:* 5%

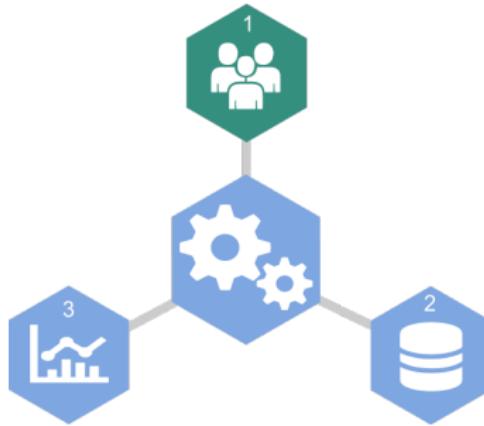
 CrowdFlower

# Machine Learning na prática

1. **Definição dos objetivos:** trata-se do primeiro passo de qualquer análise. Devemos saber onde queremos chegar!
2. **Coleta dos dados:** envolve a coleta de material que o algoritmo utilizará para gerar conhecimento;
3. **Exploração e preparação dos dados:** é exigido um trabalho adicional na preparação desses, recodificando-os de acordo com os *inputs* esperados;
4. **Formação do modelo:** depois dos dados preparados, o pesquisador já é capaz de dizer o que é possível aprender deles, e como;
5. **Avaliação dos modelos:** avaliamos a qualidade do aprendizado, não pode ser pouco (*underfitting*) nem decorar os dados (*overfitting*);
6. **Melhoria do modelo:** se necessário, podemos melhorar o desempenho do modelo através de estratégias avançadas.



# Objetivos

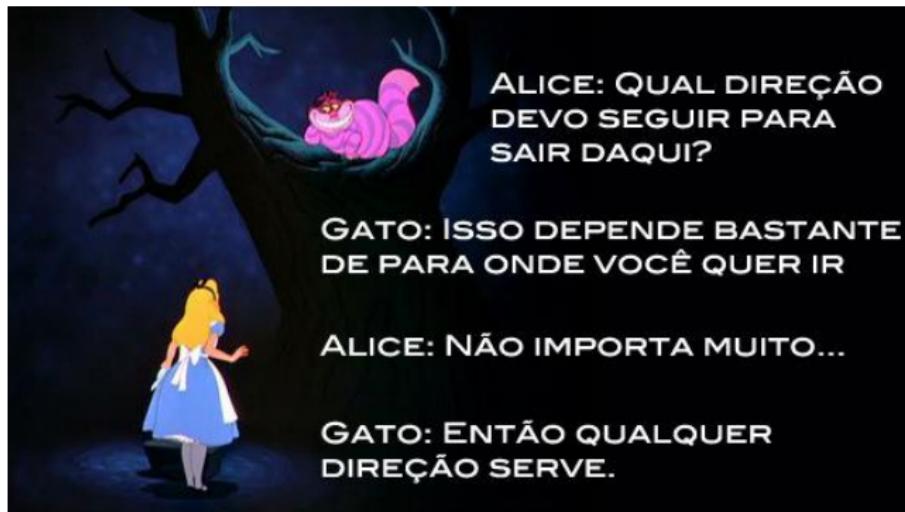


Modelos

Base de dados

# Defina seus objetivos

- ▶ Para quem não sabe onde vai, qualquer direção serve!



# Entenda o problema, depois pense como resolvê-lo

- ▶ Qual o problema na foto ao lado?
- ▶ Sendo o animal de tração, troque-o por um avião!
- ▶ Se continuar, compre um mais potente;



DSBD

# Vamos procurar por “ideias fora da caixa”

- ▶ Com o passar do tempo, criamos padrões que ficam cada vez mais estabelecidos em nossa mente;



- ▶ Este pensamento reflete em toda organização. Notamos processos funcionando da mesma maneira, meses, até anos e não fazemos nada;

**DSBD**

# Formas de pensar em cada estágio

## 1. Criativo

- ▶ Resulta em novas ideias e possibilidades;
- ▶ Sem ele, em geral, ocorre “mais do mesmo”.

## 2. Lógico positivo

- ▶ Como fazer novas ideias funcionarem;
- ▶ Sem ele mudanças não serão práticas e funcionais.

## 3. Lógico negativo (crítico)

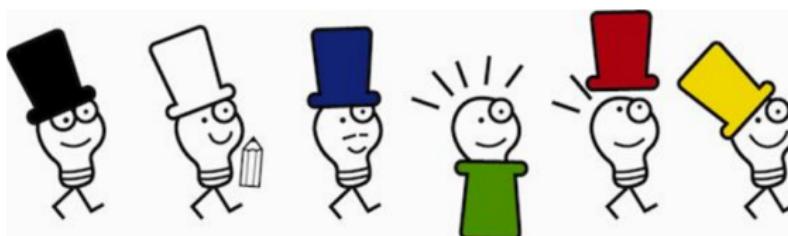
- ▶ Busca por falhas na nova ideia;
- ▶ Sem ele problemas podem não vir à tona.



# Seis chapéus do pensamento

# Seis chapéus do pensamento

- ▶ Nos ajuda a analisar um problema, uma ideia ou situação de diversas perspectivas, permitindo uma visão mais abrangente da situação;

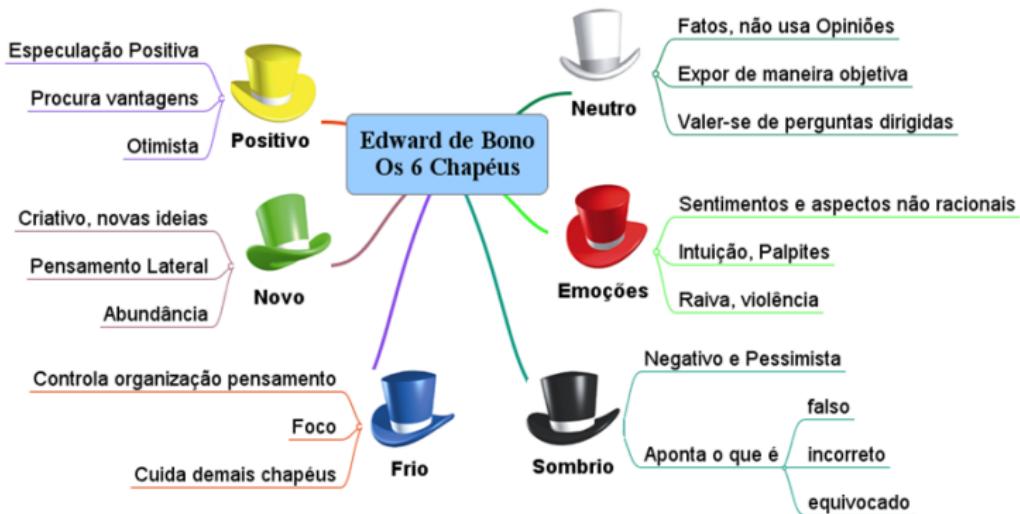


▶ Os 6 chapéus do pensamento

- ▶ De acordo com a cor do chapéu, nos focamos em apenas um aspecto do pensamento, deixando os demais de lado, até mudar do chapéu.

**DSBD**

# Seis chapéus do pensamento



ACM criado Reinaldo Geraldo jun 2012

# Técnica dos “Cinco Por quês”

# Técnica dos “Cinco Por quês”

- Foi percebido que o monumento de Abraham Lincoln deteriorava-se mais rapidamente do que qualquer outro em Washington, D.C. Por quê?



1. Porque é limpo com mais frequência que os outros monumentos. Por quê?
2. Porque tem mais dejetos de pássaros que os outros monumentos. Por quê?
3. Porque tem mais pássaros em volta deste monumento do que dos outros. Por quê?
4. Porque tem mais insetos em torno deste monumento. Por quê?
5. Porque a lâmpada que o ilumina é diferente das outras e atraí mais insetos.

# Correlações da vida real

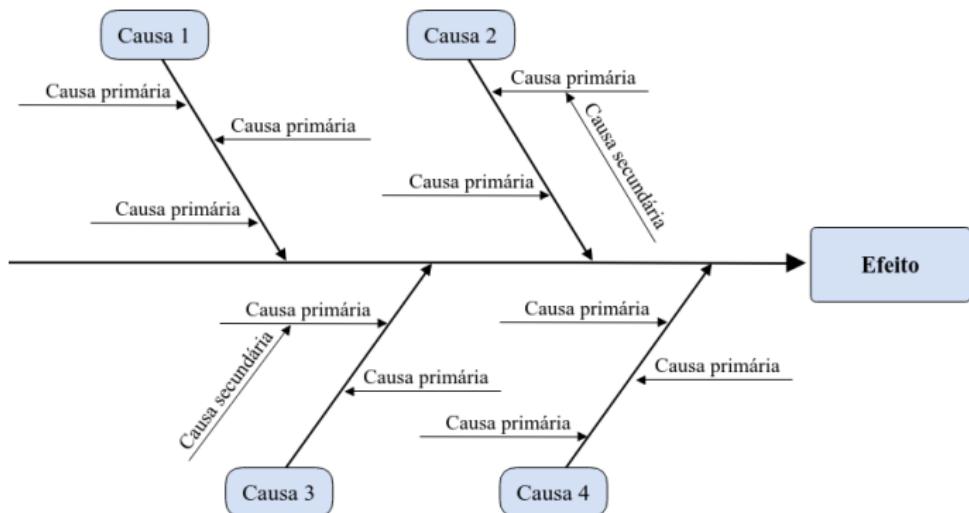
- ▶ O consumo de sorvete está correlacionado com o número de afogamentos de piscina;
- ▶ O sorvete não causa afogamentos. Ambos estão correlacionados com o clima do verão;
- ▶ Em 90% das brigas do bar que terminaram em uma morte, a pessoa que começou a briga morreu;
- ▶ Claro, é a pessoa que sobreviveu contando a história;
- ▶ Terapia de reposição hormonal está correlacionada com uma menor taxa de doença coronária;
- ▶ Pessoas que realizam reposição hormonal, geralmente, pertencem à grupos socioeconômicos mais elevados, com hábitos mais saudáveis;



# Diagrama de causa e efeito

# Diagrama de causa e efeito

- O Diagrama de causa e efeito ajuda a descobrir, organizar e resumir todo esse conhecimento atual, alinhando a equipe à respeito do problema;



# Diagrama direcionador



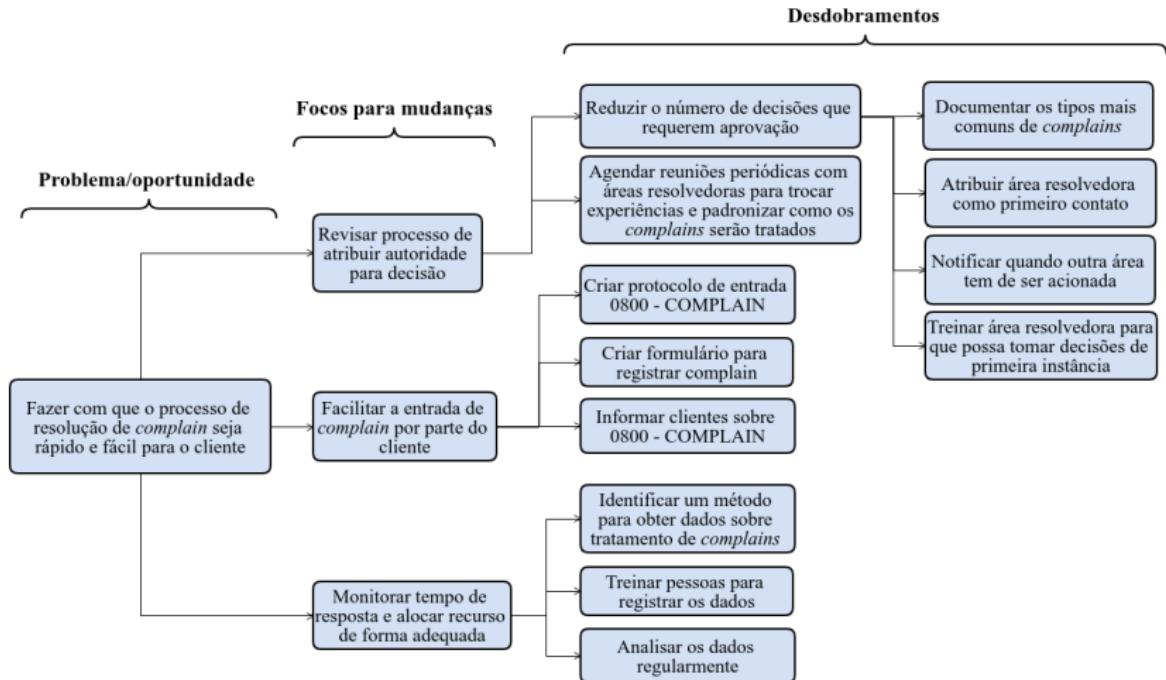
# Diagrama direcionador

- ▶ Assim como placas e faixas auxiliam no trânsito, essa técnica contribui para a busca de soluções nas diversas fases das análises.



- ▶ Como uma espécie de mapa, ele aponta caminhos ou alternativas que podem ser tomados pelo grupo de trabalho do projeto.

# Diagrama direcionador

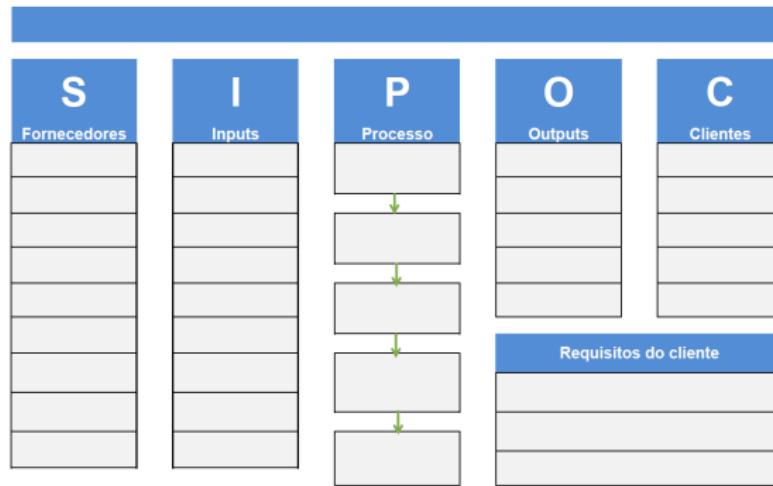


# SIPOC



# O que é?

- É uma ferramenta para representar aspectos relevantes do processo. Se não entende o que se faz, dificilmente conseguiremos melhorar;

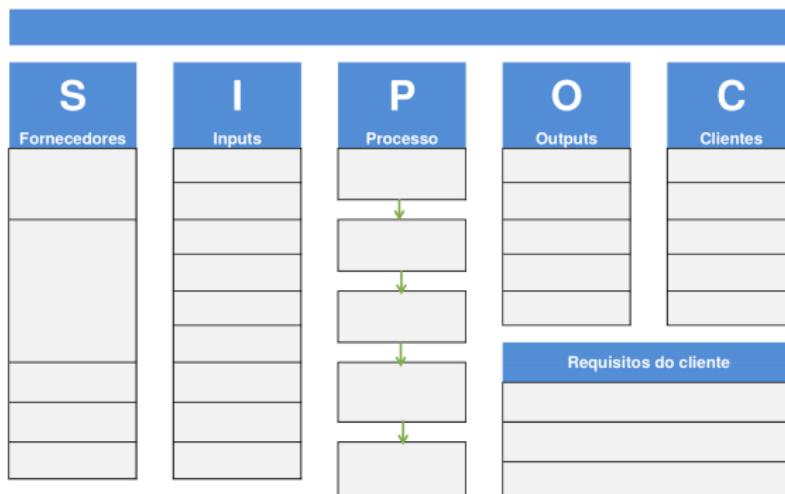


- Apresenta uma visão macro do processo, portanto não é recomendado que se inclua detalhes específicos.

**DSBD**

# Exemplo: Realizar exame de sangue

1. Definir as saídas do processo: As saídas são os resultados de um processo (ex: um relatório, uma carta, um produto etc.);



# Exemplo: Realizar exame de sangue

1. Definir as saídas do processo: As saídas são os resultados de um processo (ex: um relatório, uma carta, um produto etc.);



# Exemplo: Realizar exame de sangue

- Definir os clientes do processo: São as pessoas, empresas ou outros processos internos da organização que recebem as saídas do processo;



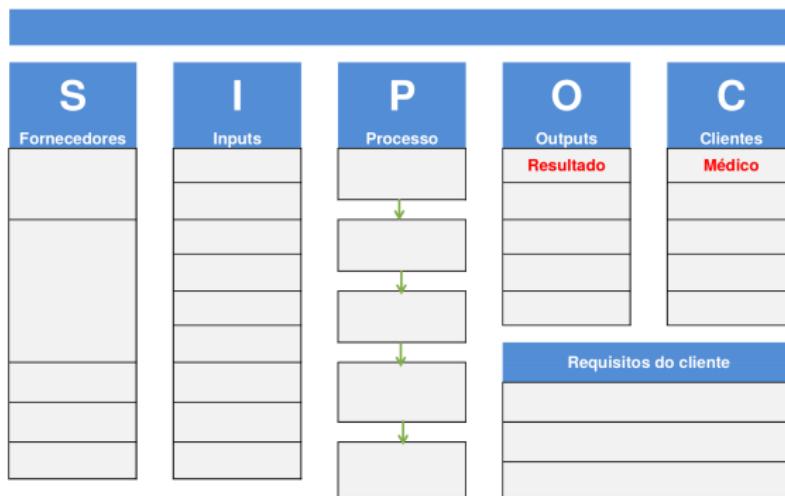
# Exemplo: Realizar exame de sangue

- Definir os clientes do processo: São as pessoas, empresas ou outros processos internos da organização que recebem as saídas do processo;



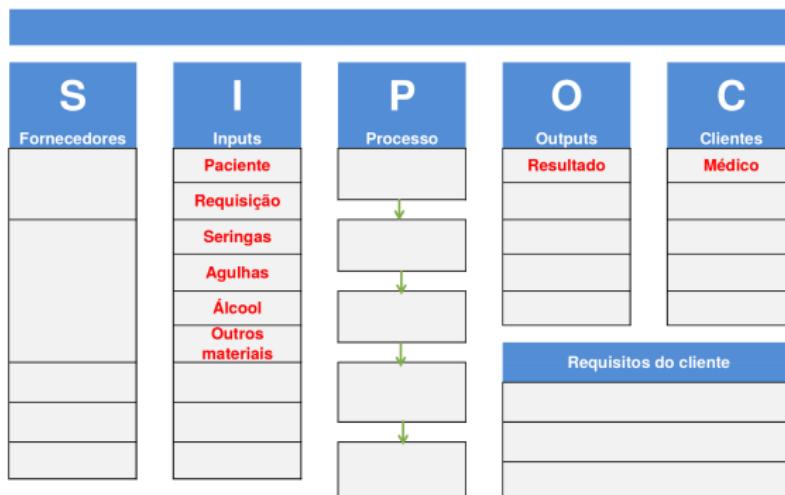
# Exemplo: Realizar exame de sangue

3. Definir as entradas do processo: Essas são os elementos necessários para iniciar o processo;



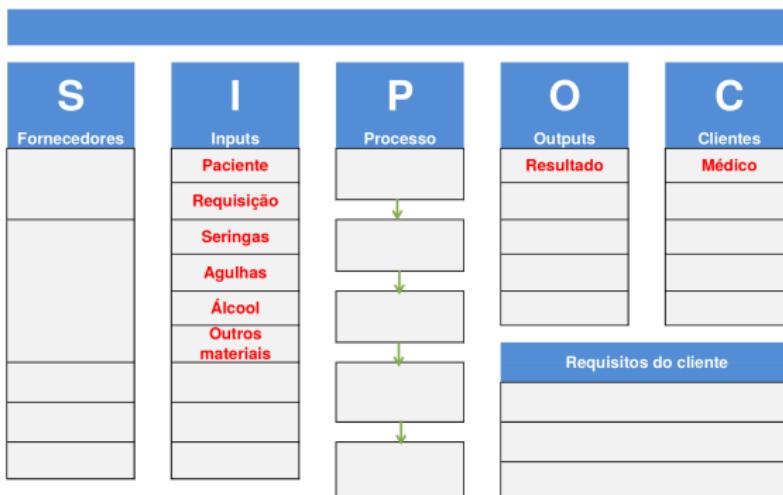
# Exemplo: Realizar exame de sangue

3. Definir as entradas do processo: Essas são os elementos necessários para iniciar o processo;



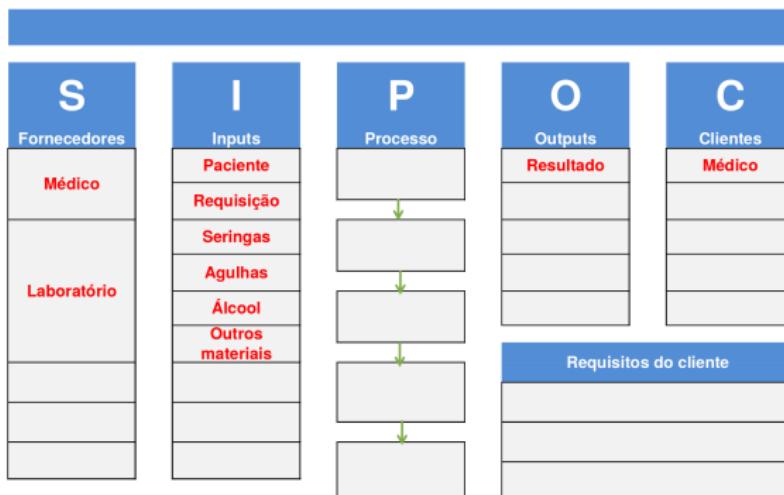
# Exemplo: Realizar exame de sangue

4. Definir os fornecedores do processo: são as pessoas ou outros processos que fornecem as entradas. Toda entrada deverá possuir um fornecedor;



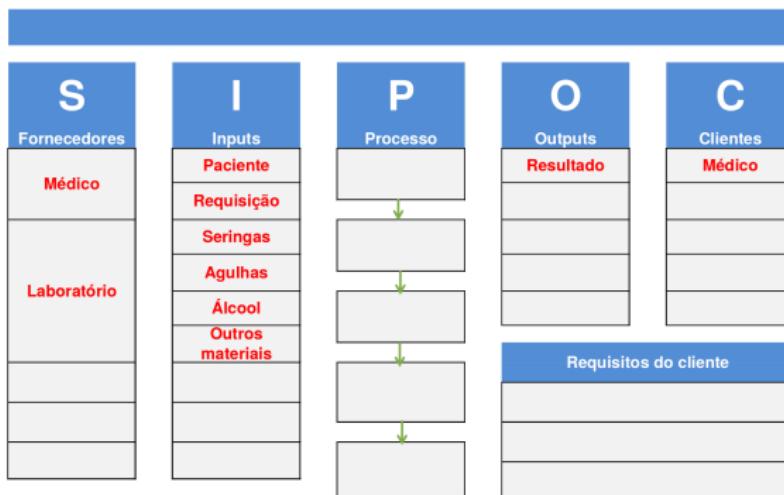
# Exemplo: Realizar exame de sangue

4. Definir os fornecedores do processo: são as pessoas ou outros processos que fornecem as entradas. Toda entrada deverá possuir um fornecedor;



# Exemplo: Realizar exame de sangue

5. Definir as macro atividades do processo: são os passos que transformam as entradas em saídas (no máximo 4 ou 5 passos).



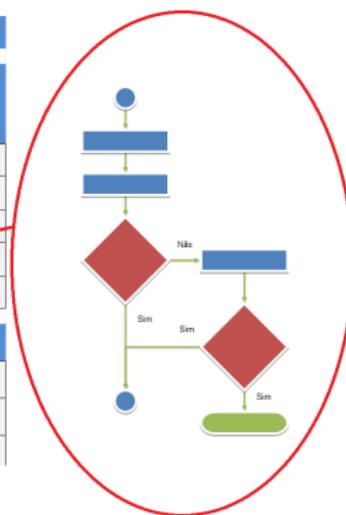
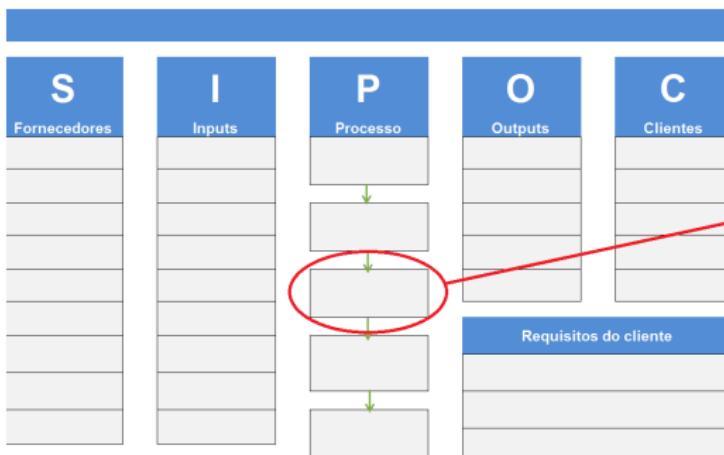
# Exemplo: Realizar exame de sangue

5. Definir as macro atividades do processo: são os passos que transformam as entradas em saídas (no máximo 4 ou 5 passos).



# Fluxograma

- No SIPOC, na etapa P, descrevemos as etapas macro do processo em 5 passos. No fluxograma detalhamos melhor cada passo.



DSBD

# Exemplo de Contrato do projeto

Patrocinador: Michellangelo

Líder da Equipe: Donatello

Demais integrantes: Leonardo, Raphael

Contexto/descrição: Nossa empresa possui 42 máquinas que precisam realizar setup para trocar o molde de injeção.

Problema: o tempo de máquina parada é considerado excessivamente alto.

Q1. O que estamos tentando realizar?	Q2. Como saberemos que a mudança é uma melhoria?		
Objetivos (O que, onde, quanto, quando)	Indicadores	Desempenho atual	Meta
Reducir o tempo de máquina parada para setup, na linha de injeção plástica, em 50% até janeiro de 2017.	Tempo (em horas por semana) de máquina parada para setup	500 horas semanais (média)	250 horas semanais

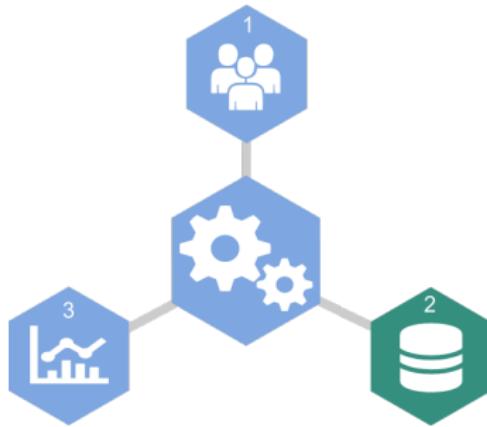
Business case: Com um custo de R\$140,00 por hora de máquina parada, a redução no tempo de máquina parada de 500 para 250 horas semanais reduzirá os custos em torno de R\$35.000,00 por semana ou R\$140.000,00 por mês.

Q3. Atividades iniciais do projeto: (1) Preparar um SIPOC do processo de troca de molde (2) Coletar dados de tempo de parada estratificados (aguardando liberação, aguardando equipe de setup, aguardando OS).

Restrições para as atividades: Não será possível qualquer investimento (máquinas, contratações, etc.).



# Objetivos



Modelos

Base de dados

# The Art of Feature Engineering

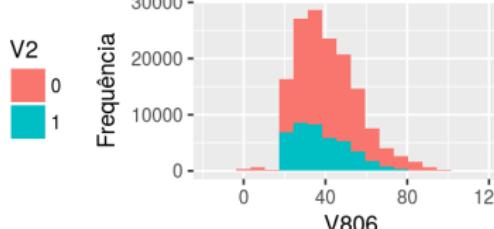
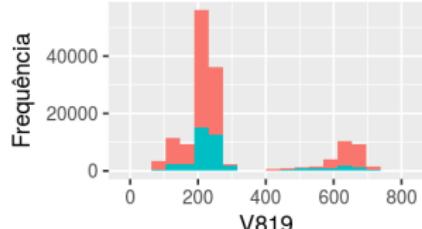
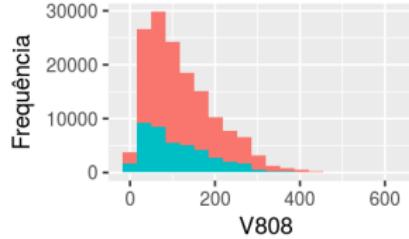
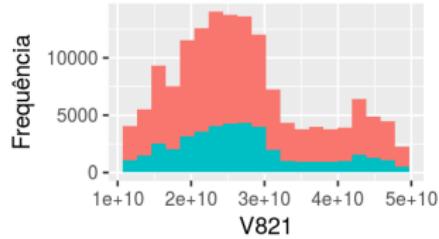
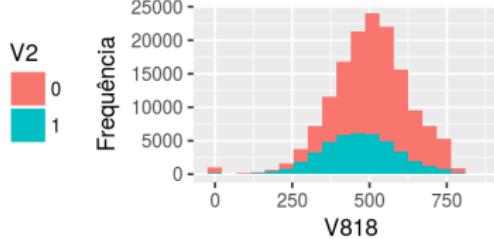
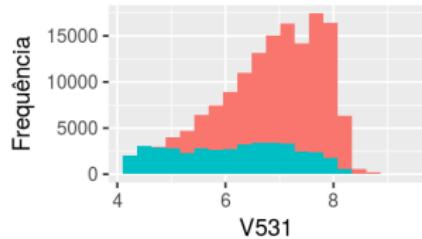


- ▶ **Feature engineering** é a arte de extrair informação dos dados já obtidos:
  1. Criação de características;
  2. Dados faltantes;
  3. Dados desbalanceados;
  4. Variáveis correlacionadas.

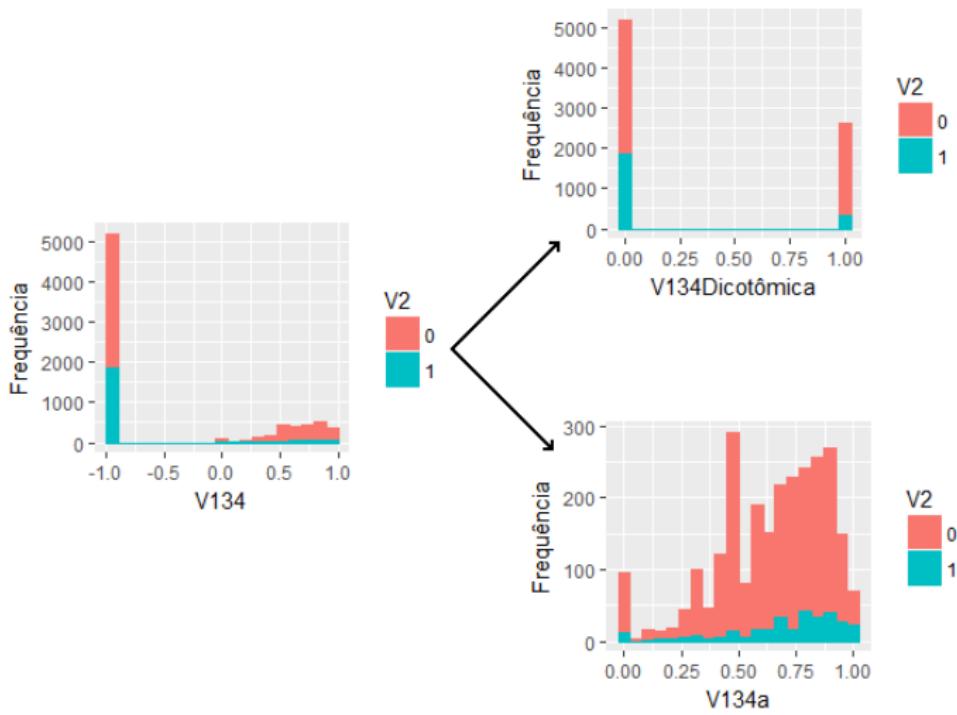
# Criação de características



# Exemplo

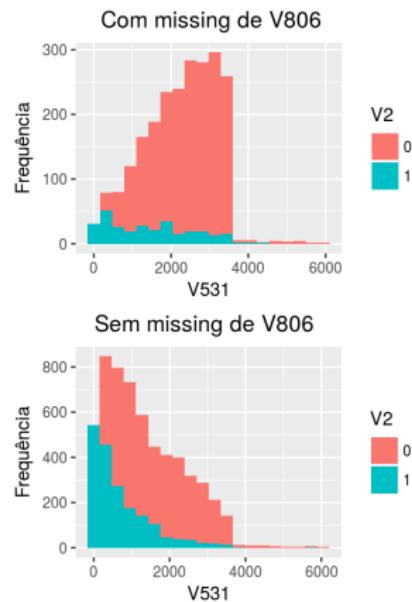
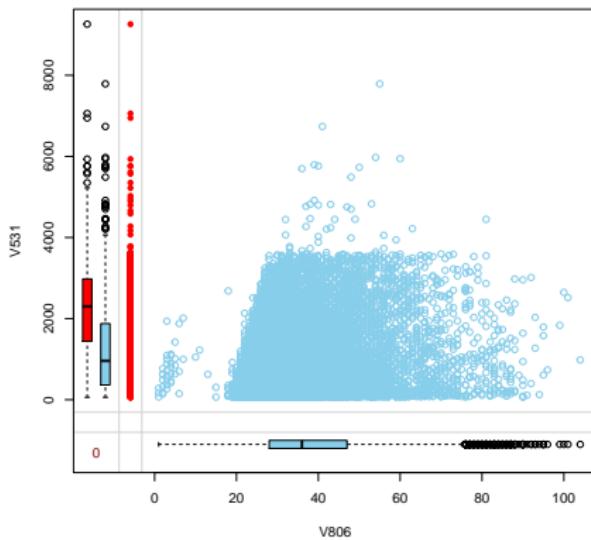


# Exemplo



# Dados faltantes

# Exemplo



# Seja cauteloso na imputação dos dados!



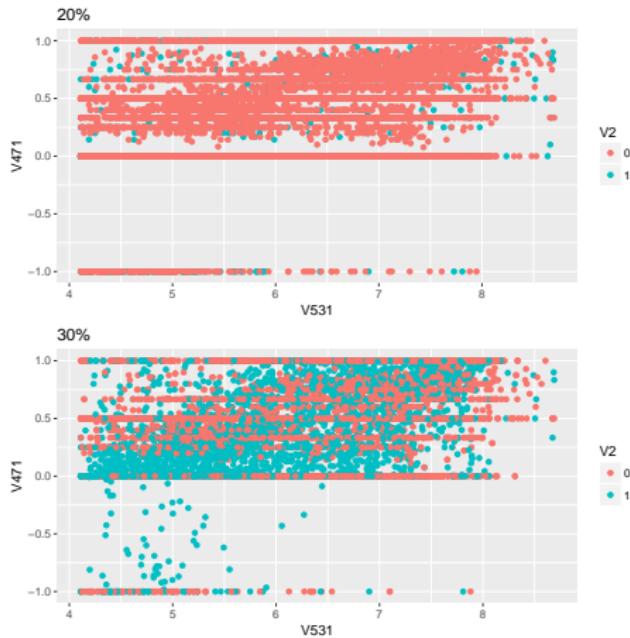
“The idea of imputation is both seductive and dangerous.” **D.B. Rubin**

# Dados desbalanceados



# Dados desbalanceados

1. **SMOTE (Synthetic Minority Over-sampling Technique):** remostrar o conjunto de dados original por superamostragem da classe minoritária;



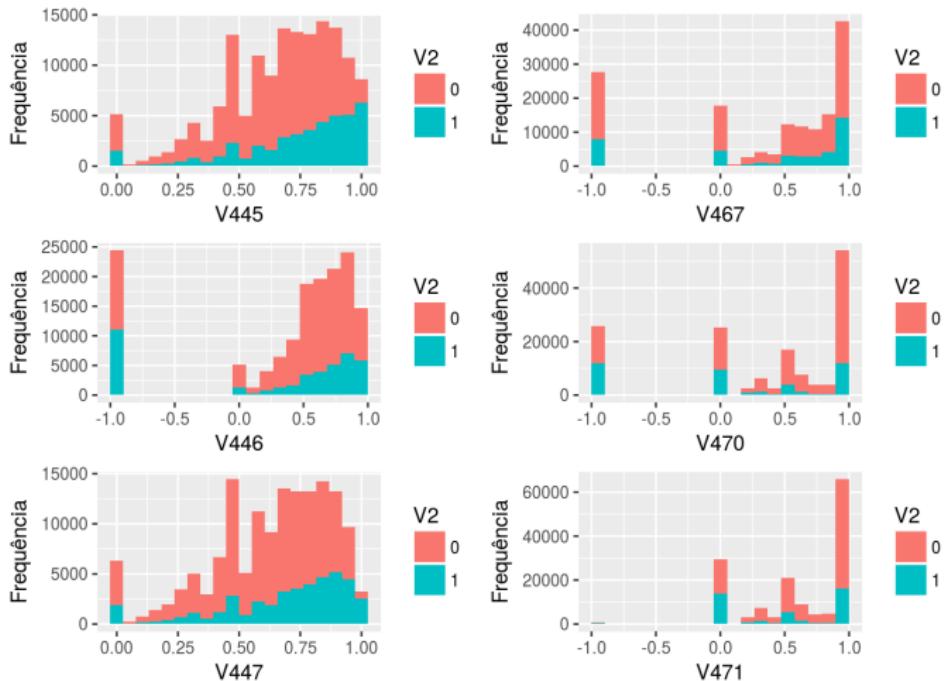
# Dados desbalanceados

2. **Utilizar diferentes algoritmos:** abordagens simples como Árvores, geralmente, apresentam um bom desempenho em dados desbalanceados;
3. **Modelos penalizados:** existem várias versões de algoritmos penalizados como *penalized-SVM* e *penalized-LDA*.
4. **Conceitos em outras perspectivas:** há vários campos dedicados a dados desbalanceados. P. ex., [detecção de anomalias](#), [detecção de alterações](#);
5. **Ser criativo:** busque inspirações, por exemplo, em respostas do Quora: “[in classification, how do you handle an unbalanced training set?](#)”
  - ▶ “Decomponha a classe maior em pequenas outras classes”.
  - ▶ “Reamostre os dados desbalanceados em não somente um conjunto balanceado, mas vários”.



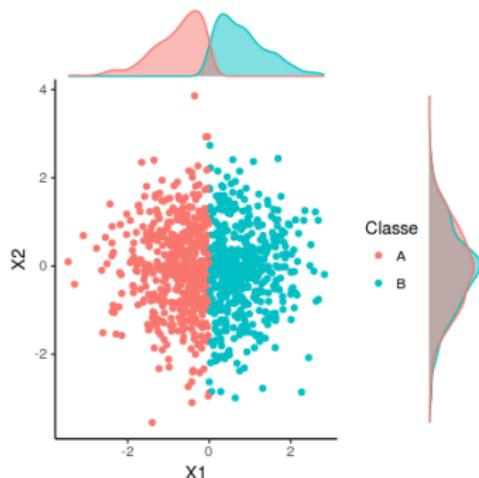
# Variáveis correlacionadas

# Variáveis correlacionadas



# Ranking de características individuais

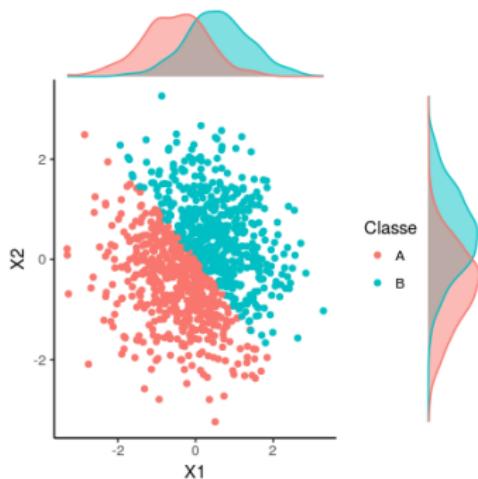
- Neste caso, a variável  $x_1$  é relevante individualmente, e  $x_2$  não ajuda a obter uma melhor separação.



- O ranking de características individuais funciona bem. A característica que proporciona uma boa separação de classe será escolhida.

# Rotações no espaço de características

- ▶ A figura anterior foi obtida da figura abaixo após rotação de 45 graus. Agora, para alcançar a mesma separação, são necessárias  $x_1$  e  $x_2$ .

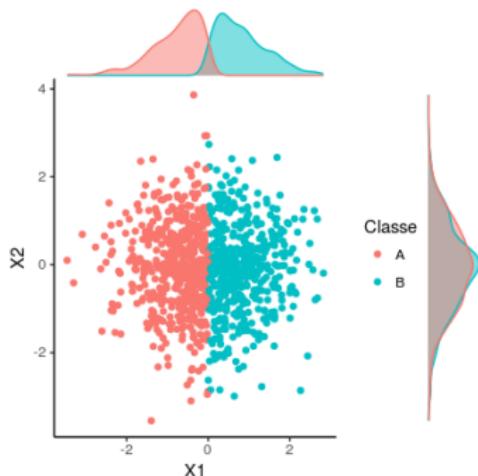


- ▶ Vários métodos de pré-processamento, como a análise de componentes principais (PCA), realizam transformações lineares (como a rotação).

DSBD

# Rotações no espaço de características

- **Pergunta:** no caso abaixo, você manteria  $x_1$  e  $x_2$  ou eliminaria  $x_2$ ? Note que a noção de relevância está relacionada ao objetivo perseguido;

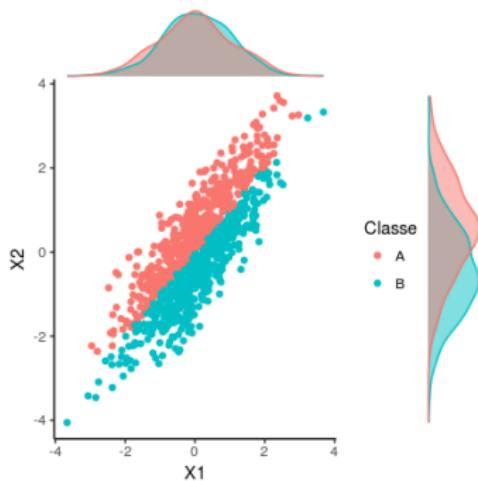


- $P(Y|X)$  não é independente de  $x_2$ , mas a taxa de erro do classificador Bayes ideal é a mesma se  $x_2$  é mantida ou descartada.

DSBD

# Características individualmente irrelevantes

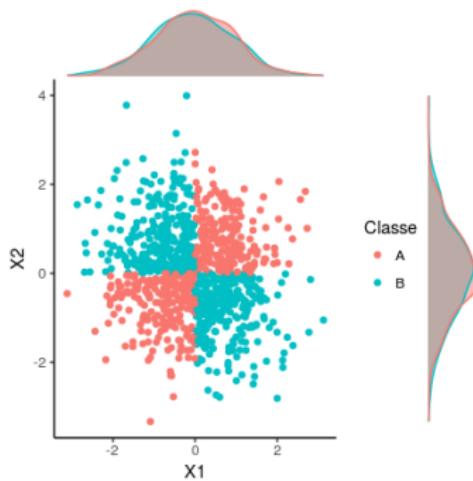
- ▶ Nota-se uma separação linear, em que as características individualmente irrelevantes ajudam a obter uma melhor separação quando em conjunto;



- ▶ Neste caso, é justificável o uso de métodos multivariados, que utilizam o poder preditivo das características em conjunto.

# Características individualmente irrelevantes

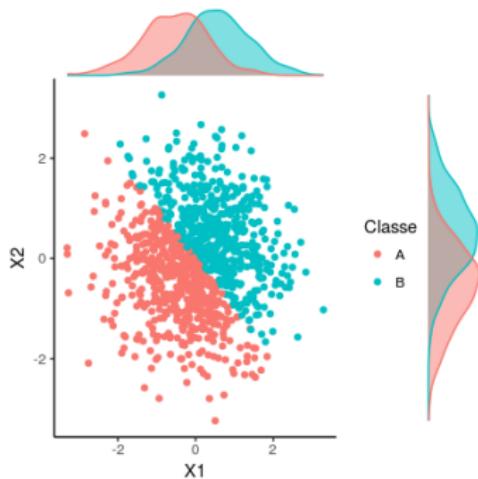
- ▶ O que podemos dizer sobre o problema abaixo, alguma característica é individualmente irrelevante?



- ▶ Este caso é conhecido como problema do tabuleiro de xadrez. As características são conjuntamente relevantes.

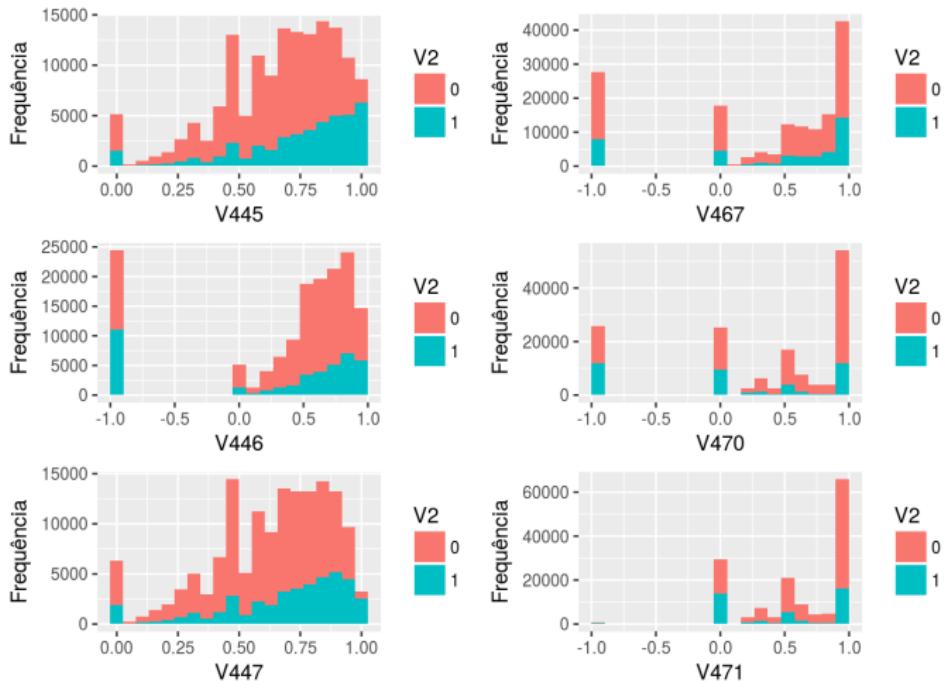
# Características aparentemente redundantes

- ▶ A redução do ruído pode ser alcançada quando características com distribuições projetadas idênticas.



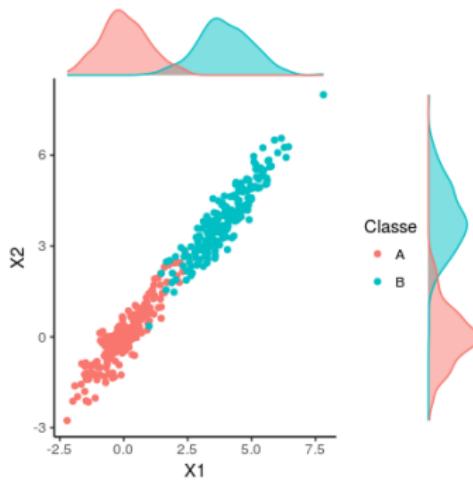
- ▶ A distribuição bidimensional mostra uma separação de classe melhor, quando comparado com qualquer característica individual.

# Variáveis correlacionadas



# Correlação não implica redundância

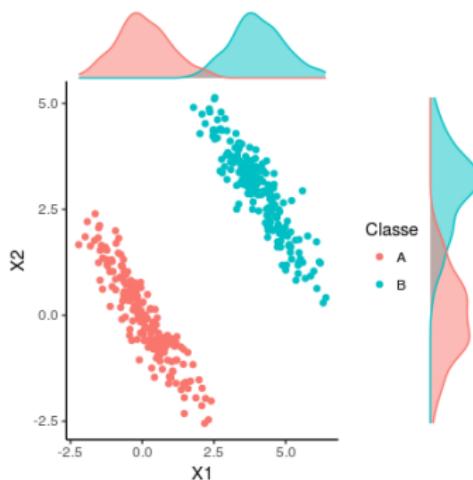
- ▶ Geralmente, se pensa que correlação de característica significa redundância de recurso.



- ▶ As características são redundantes. I. e., a separação de classe não é aprimorada ao se considerar as variáveis conjuntamente.

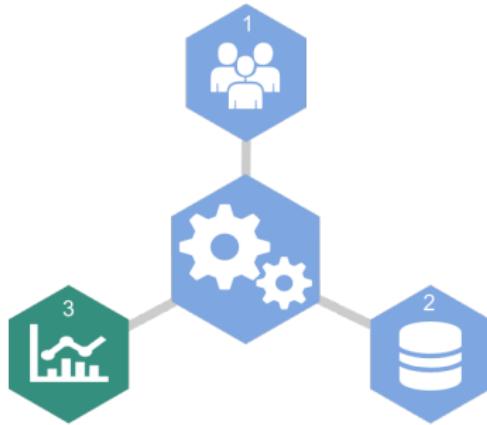
# Correlação não implica redundância

- ▶ Geralmente, se pensa que correlação de característica significa redundância de recurso.



- ▶ Apesar das projeções serem semelhantes à anterior (e correlacionadas), elas não são redundantes.

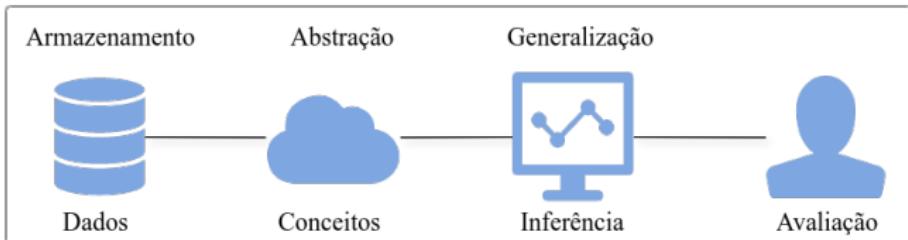
# Objetivos



Modelos

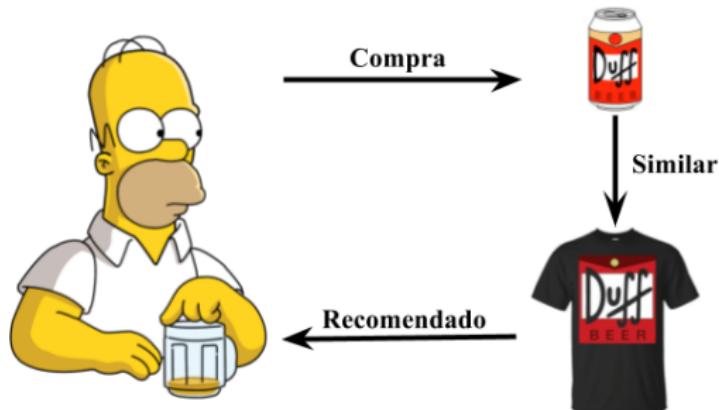
Base de dados

# Como as máquinas aprendem?



- ▶ **Armazenamento dos dados:** utiliza a observação para fornecer uma base para o raciocínio adicional;
- ▶ **Abstração:** envolve a tradução dos dados em representações e conceitos;
- ▶ **Generalização:** cria conhecimento e inferência que direcionam ações em novos contextos;
- ▶ **Avaliação:** fornece um mecanismo de *feedback* para medir a utilidade do conhecimento adquirido e informar potenciais melhorias.

# Exemplo



# Todos os aspectos são importantes



Dados bons



+



Modelo ruim



=



Resultado ruim



Dados ruins



+



Modelo perfeito



=

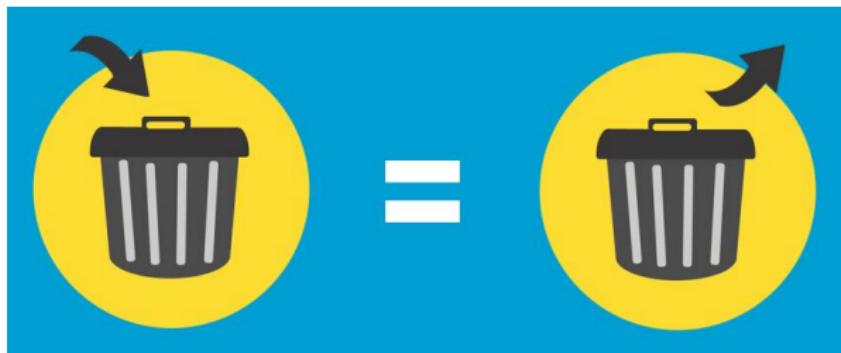


Resultado ruim



# Os limites do Machine Learning

- ▶ Machine Learning tem pouca flexibilidade para extrapolar os parâmetros de aprendizagem e não conhece o senso comum!



- ▶ Ele é tão bom quanto os dados são para ensinar. É um paradigma “Garbage in, garbage out!”

# Interpretabilidade vs Flexibilidade

