



MACHINE LEARNING  
E DATA MINING



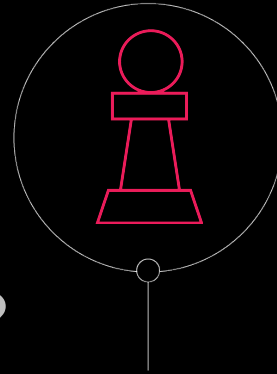
# BIG DATA SCIENCE

## MAIS SOBRE MODELOS REGRESSIVOS: TRANSFORMAÇÕES... – PARTE 3

CLASSIFICAÇÃO E CLUSTERIZAÇÃO  
TEORIA DAS PROBABILIDADES E CONCEITO DE SIMILARIDADE

DIÓGENES JUSTO

# O QUE É BIG DATA SCIENCE?



Existem por volta de  
169.518.829.100.544.000.000.000.000.000 formas  
de jogar os 10 primeiros movimentos de xadrez



# MÉTODOS DOS MÍNIMOS QUADRADOS

## DEMONSTRAÇÃO MATEMÁTICA


<http://isites.harvard.edu/fs/docs/icb.topic515975.files/OLSDerivation.pdf>

## EM TRANSACT-SQL (SQL SERVER)

<https://ayadshammout.com/2013/11/30/t-sql-linear-regression-function/>

## EM JAVA

<https://introcs.cs.princeton.edu/java/97data/LinearRegression.java.html>




A photograph of Michael Stonebraker, a man with grey hair and glasses, wearing a red polo shirt. He is standing and gesturing with his right hand, pointing upwards. He is positioned in front of a large projection screen.

### Data Scientist Job Description

- Find and clean the data (vast majority of his time right now -- forward pointer to Variety)
- Until (tired) {  
    Data management operation(s);  
    Complex analytics operations(s);  
}

DBg Database Group  
MIT Computer Science and Artificial Intelligence Lab 11



Michael Stonebraker | Big Data is (at least) Four Different Problems



# REVISÃO DO CONTEÚDO

## BOXPLOT E HISTOGRAMA

```
install.packages("UsingR")  
library(UsingR)  
simple.hist.and.boxplot(mtcars$mpg)
```

**HISTOGRAMA**

**DISTRIBUIÇÃO**

<http://shabal.in/visuals/histogram2density.gif>



## REGRESSÃO LINEAR SIMPLES

Excel: Dados, Análise de Dados, Regressão Linear

R:  $\text{lm}(y \sim x)$

## DUMMIES DE SAZONALIDADE

Excel: Uma variável para cada evento sazonal, atribuindo 1 para sazonalidade e 0 caso contrário.

R:  $\text{lm}(y \sim x_1 + x_2 + \text{saz})$  onde saz é uma variável categórica para cada sazonalidade

## REGRESSÃO LINEAR MÚLTIPLA

$\text{lm}(y \sim x_1 + x_2)$

Dummy para outlier: Similar a Dummy de sazonalidade

## MODELO AUTOREGRESSIVO

Excel: Trabalhar com colunas  $y_i$ ,  $y_{i-1}$ , etc.



# UM POUCO MAIS SOBRE REGRESSÃO

# REGRESSÃO LINEAR

Pontos de Atenção

- MUITAS VARIÁVEIS INDEPENDENTES:
  - Modelo muito grande
  - Multicolinearidade: variáveis com alta correlação entre si:  
<http://blog.minitab.com/blog/adventures-in-statistics-2/what-are-the-effects-of-multicollinearity-and-when-can-i-ignore-them>



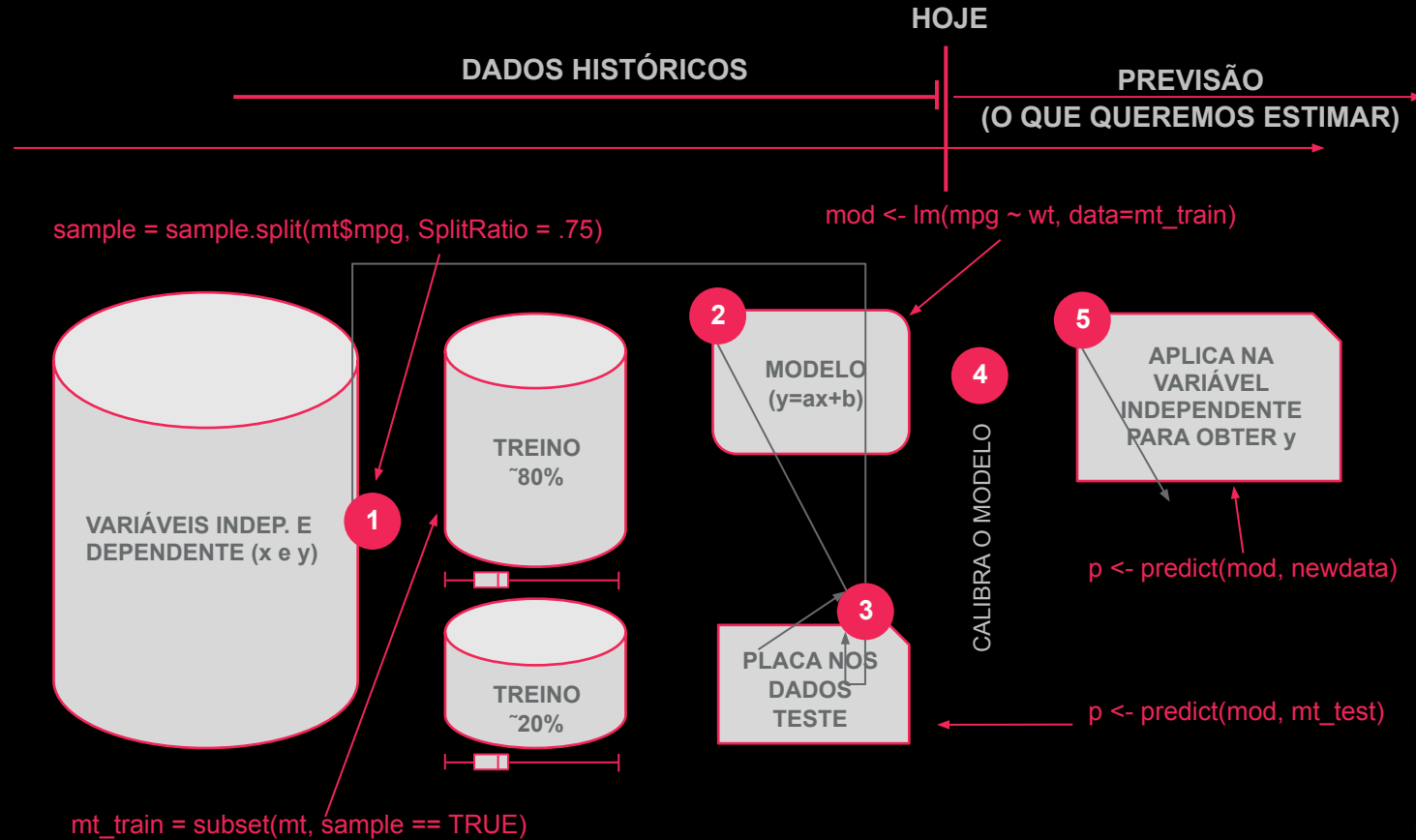
PREDICTING

Após elaborarmos um modelo  
(i.E., Calcularmos  $\alpha_1$ ,  $\alpha_2$  ...) montamos  
a equação e aplicamos os novos valores  
(variáveis exógenas) para calcular os valores  
da variável dependente (endógena)

# O R POSSUI UMA FUNÇÃO PARA ESTA TAREFA: PREDICT

```
library(caTools)
mt <- mtcars
set.seed(88)
sample = sample.split(mt$mpg, SplitRatio = .75) mt_train
= subset(mt, sample == TRUE)
mt_test = subset(mt, sample == FALSE)
mod <- lm(mpg ~ wt, data=mt_train)
p <- predict(mod, mt_test) plot(mt$mpg ~
mt$wt) points(mt_test$mpg~ mt_test$wt,
pch=7) points(p~mt_test$wt, pch=8)
```

## QUAIS DADOS USAR NA ESTIMAÇÃO DO MODELO?



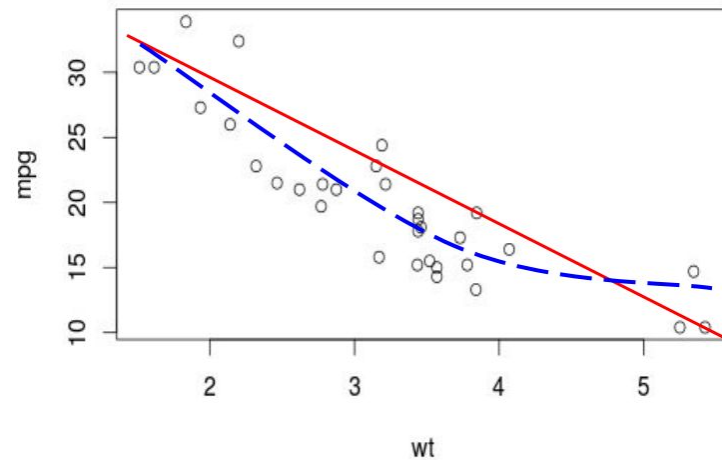


# TRANSFORMAÇÕES



# REGRESSÃO LINEAR

Parece adequada a abordagem linear para este tipo de problema?



# ESCALA LOG X LIN

1 2 3 4 5 6 7 8 9 10



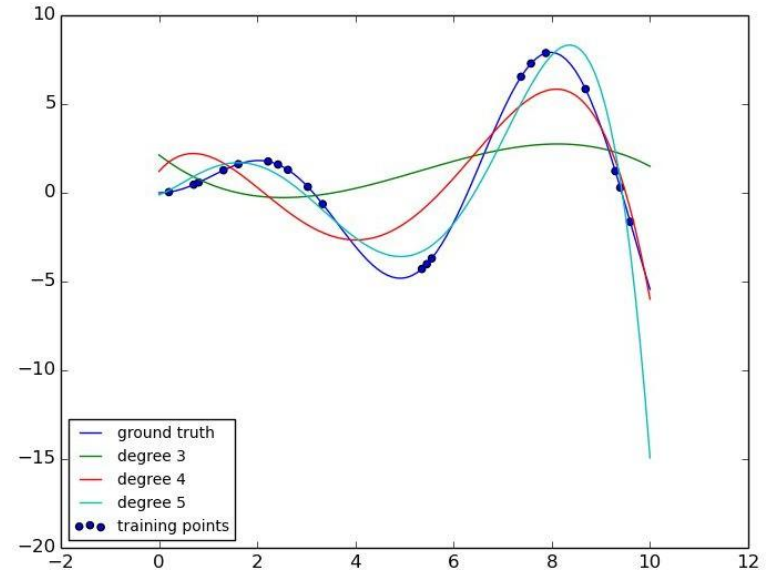
1 2 3 4 5 6 7 8 9 10



# TRANSFORMAÇÕES EM MODELOS LINEARES

Método	Transformação(s)	Equação de regressão	Valores estimados ( $\hat{y}$ )
Regressão Linear Simples	Nenhum	$y = b_0 + b_1x$	$\hat{y} = b_0 + b_1x$
Modelo exponencial	Variável dependente = $\log(y)$	$\log(y) = b_0 + b_1x$	$\hat{y} = 10^{b_0 + b_1x}$
Modelo quadrático	Variável dependente = $\sqrt{y}$	$\sqrt{y} = b_0 + b_1x$	$\hat{y} = (b_0 + b_1x)^2$
Modelo recíproco	Variável dependente = $1/y$	$1/y = b_0 + b_1x$	$\hat{y} = 1 / (b_0 + b_1x)$
Modelo logarítmico	Variável dependente = $\log(x)$	$y = b_0 + b_1\log(x)$	$\hat{y} = b_0 + b_1\log(x)$
Modelo potência	Variável dependente = $\log(y)$ Variável dependente = $\log(x)$	$\log(y) = b_0 + b_1\log(x)$	$\log(y) = 10^{b_0 + b_1\log(x)}$

# TRANSFORMAÇÕES EM MODELOS LINEARES: polinomial



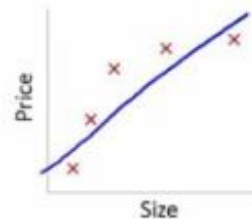
# COMPARANDO MODELOS LINEARES

```
m <- mtcars
set.seed(33)
m <- m[sample(nrow(m)),]
mod <- lm(mpg~wt, data=mtr)
#mod <- lm(mpg~poly(wt,2), data=mtr)
#mod <- lm(mpg~poly(wt,3), data=mtr)
#mod <- lm(mpg~poly(wt,4), data=mtr)
#mod <- lm(mpg~poly(wt,5), data=mtr)
#mod <- lm(mpg~poly(wt,6), data=mtr)
#mod <- lm(mpg~wt+hp, data=mtr)
#mod <- lm(mpg~wt+disp, data=mtr)
#mod <- lm(mpg~wt+cyl, data=mtr)
#mod <- lm(mpg~wt+as.factor(cyl), data=mtr)
```

```
sum(mod$residuals^2)      ← Erro em Treino (fitting)
summary(mod)$r.squared    ← x
prev <- predict(mod, newdata=mts)  R^2
e <- mts$mpg - prev       ← x
sum(e^2)                  ← Erro em Teste (validação)
```

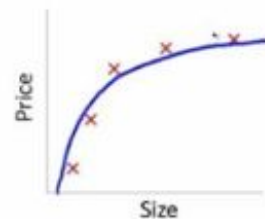
# OVERFIT

Fonte: Prof. Andrew  
Ng



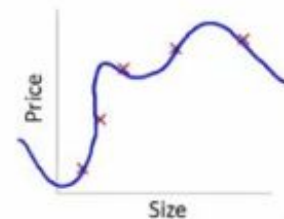
$$\theta_0 + \theta_1 x$$

High bias  
(underfit)



$$\theta_0 + \theta_1 x + \theta_2 x^2$$

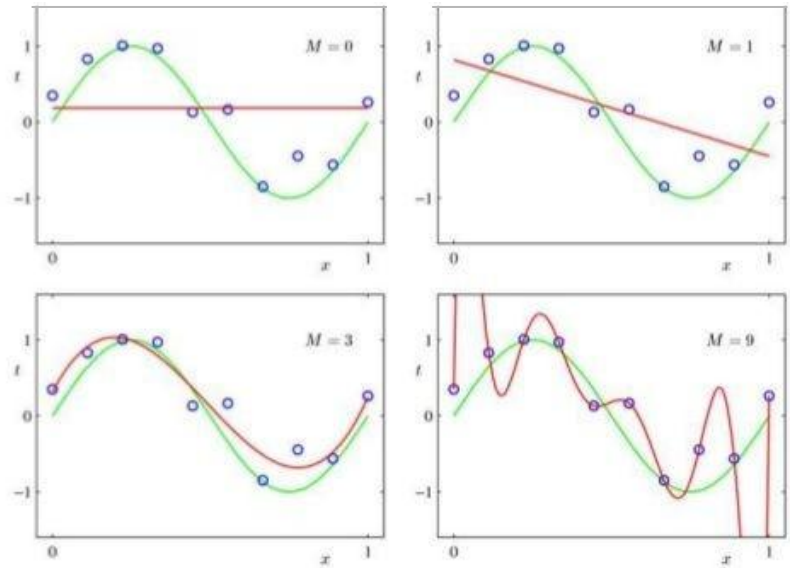
"Just right"



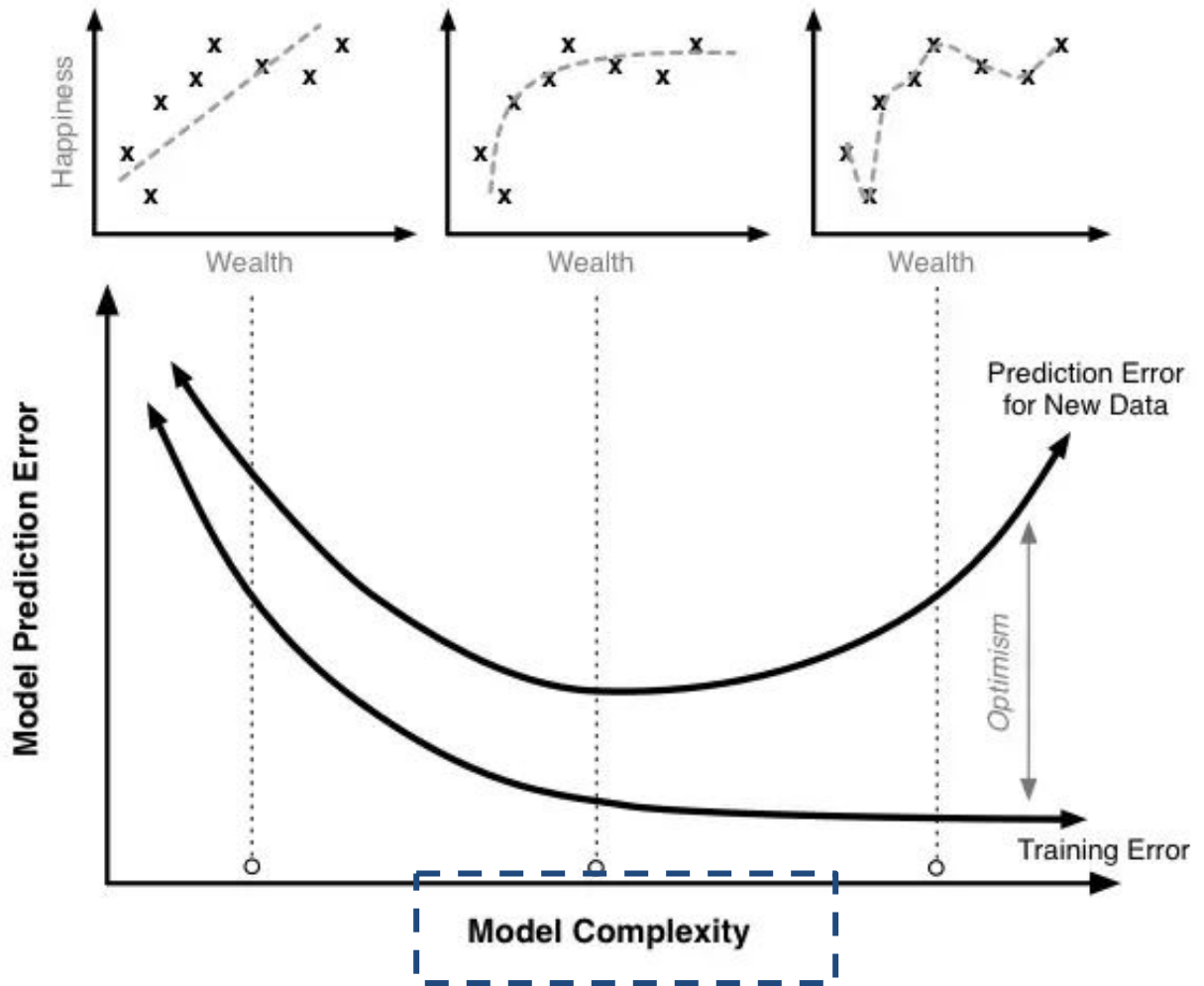
$$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

High variance  
(overfit)

# OVERFIT



# OVERFIT





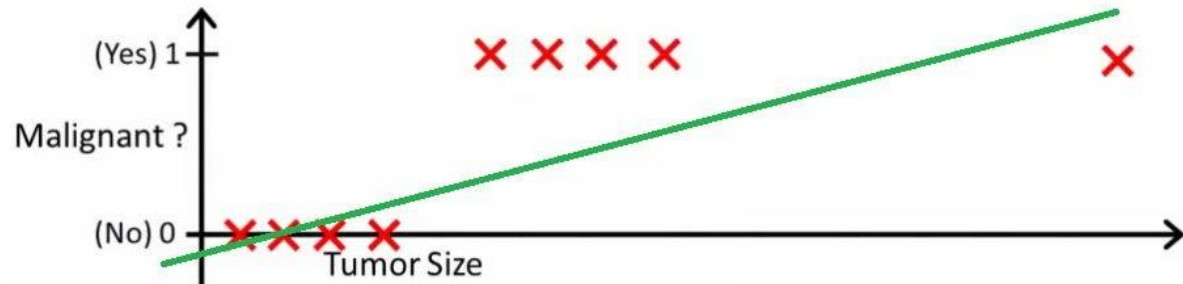
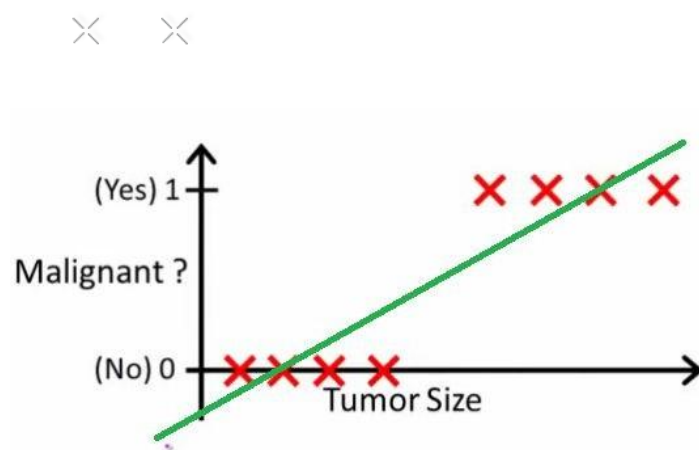


# CLASSIFICAÇÃO

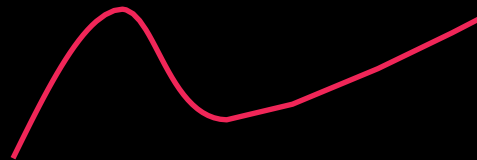
# REGRESSÃO x CLASSIFICAÇÃO

Parece adequada a abordagem linear  
para este tipo de problema?

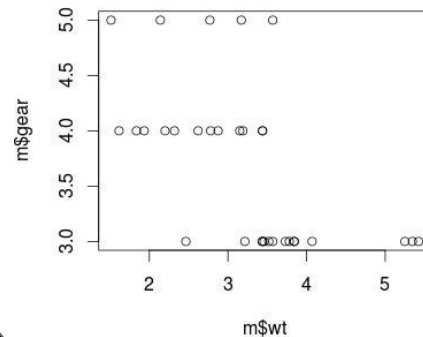
# REGRESSÃO X CLASSIFICAÇÃO



REGRESSÃO (EM GERAL):  
variável endógena contínua



CLASSIFICAÇÃO (EM GERAL):  
variável endógena discreta



# STEPS IN DATA ANALYSIS

- Define the **QUESTION**
- Define the **IDEAL DATA SET**
- Determine **WHAT DATA** you can access
- **OBTAIN** the data
- **CLEAN** the data
- Exploratory **DATA ANALYSIS**
- Statistical **PREDICTION/MODELING**
- **INTERPRET** Results
- **CHALLENGE** results
- Synthesize/**WRITE UP** results
- Create reproducible **CODE**



# PROBABILIDADES

A teoria das probabilidades foi formalizada a partir de estudos contemporâneos dos matemáticos Fermat e Pascal (1654) entre outros. No entanto, estima-se que a observação empírica se dê há séculos ou milênios, desde que o homem criou os jogos chamados "de azar".

Calculamos a probabilidade de um evento ocorrer a partir da análise das maneiras possíveis dele ocorrer e do número total de resultados possíveis.

$P=x/t$  , onde  $x$  são as maneiras possíveis de um evento ocorrer e  $t$  o número total de resultados possíveis.

## PROBABILIDADE CLÁSSICA A PRIORI

Análise do evento e das maneiras dele ocorrer.

Por exemplo, jogar um dado esperar tirar o número 5. O 5 é a maneira esperada, somente presente em 1 das 6 faces possíveis do dado cair. Portanto

$$P_5 = 1/6$$

## PROBABILIDADE CLÁSSICA EMPÍRICA

Realiza-se a medição de determinado evento para determinar a probabilidade. Por exemplo, número de falhas de uma máquina (up-time).



# CALCULANDO R

# Mulheres/Total de passageiros: muitas alternativas...

```
dsTRAIN <- read.csv(file.choose())  
d <- dsTRAIN  
(1) NROW(d[d$Sex=='female',])/NROW(d)
```

```
(2) install.packages("sqldf")  
    (a) library(sqldf)  
    (b) sqldf('select Sex,count(Survived)  
              from d group by Sex')
```

```
(3) install.packages("plyr")  
    (a) library(plyr)  
    (b) count(d, c("Sex"))
```




## EVENTO

cada resultado possível de uma variável.

## EVENTO COMBINADO

evento que possui duas ou mais características.

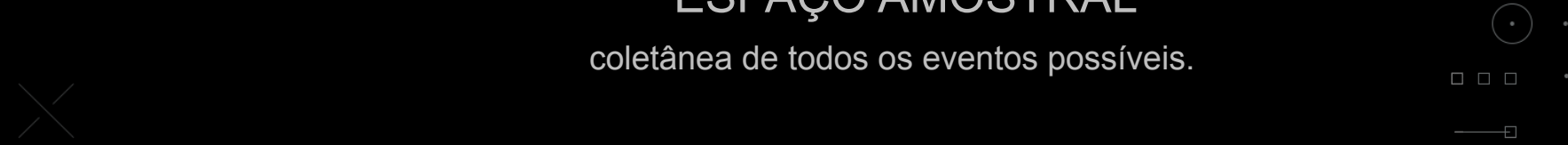


## COMPLEMENTO

$A'$ =complemento de A. Todos os eventos que não o próprio, dentro da amostra.

## ESPAÇO AMOSTRAL

coletânea de todos os eventos possíveis.



# PROBABILIDADES

## Regras Básicas

- Uma probabilidade é UM NÚMERO ENTRE 0 E 1
- A probabilidade de um evento SOMADA A SEU COMPLEMENTO É 1
  - ex.: no caso de um dado  $P_i + P'_i = 1$
- O SOMATÓRIO de todas as probabilidades (probabilidades de todos os eventos) é 1.
  - ex.: no caso de um dado  $P_1 = P_2 = P_3 = P_4 = P_5 = P_6 = \frac{1}{6}$ ,  $\square P_i = \frac{1}{6} + \dots + \frac{1}{6} = 1$

# PROBABILIDADES

## EXEMPLO TITANIC

1 - Entre os sobreviventes, qual a probabilidade que seja mulher?

```
R.: filtra Survived=1;  
#Mulheres/Total sobreviventes  
count(d[d$Survived==1,],"Sex")
```

2 - Entre os passageiros da primeira classe do sexo feminino, qual a probabilidade de sobrevivência?

```
R.: filtra Pclass=1 e Sex='female';#Survived/  
[Totalpassageiros primeira classe mulheres]  
count(d[d$Pclass==1 &  
d$Sex=='female'], "Survived")
```




## EVENTO

cada resultado possível de uma variável.

## EVENTO COMBINADO

evento que possui duas ou mais características.

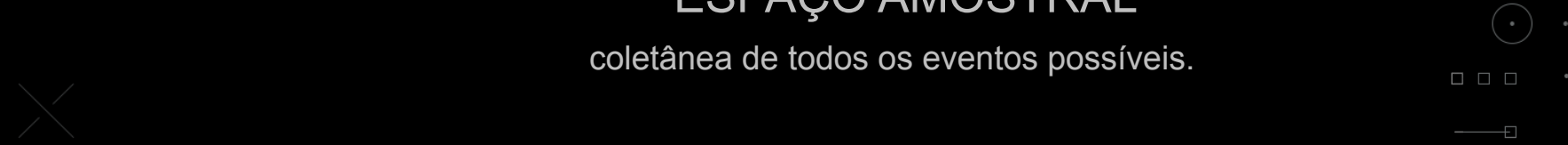


## COMPLEMENTO

$A'$ =complemento de A. Todos os eventos que não o próprio, dentro da amostra.

## ESPAÇO AMOSTRAL

coletânea de todos os eventos possíveis.



# CÁLCULOS

## REGRA GERAL DA ADIÇÃO

$P(A \text{ ou } B) \text{ ou } P(A \cup B)$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

# REGRA GERAL DA MULTIPLICAÇÃO

$P(A \text{ e } B)$  ou  $P(A \cap B)$

$$P(A \cap B) = P(A|B) * P(B)$$

## PROBABILIDADE CONDICIONAL

$$P(A|B) = P(A \cap B) / P(B)$$

Obs: Propriedade (pelo Teorema de Bayes)

$$P(B|A) = [ P(A|B) * P(B) ] / P(A)$$

# PARADOXO DE RHINE

Um experimento conduzido por David Rhine (parapsicólogo), nos anos 50, procurava avaliar a capacidade extra-sensorial.

HIPÓTESE: aquele que conseguir adivinhar as 10 cores tem percepção extra-sensoriais. Algo como: 10 mil pessoas tentaram adivinhar a cor do naipe de dez cartas.

CONCLUSÃO: 0,1% da população tem poderes extra-sensoriais.



Após a conclusão da primeira etapa, Rhine chama as pessoas que acertaram todas as cartas, parabeniza-os pois são seres com poderes especiais e refaz o experimento. Para a surpresa de Rhine, nenhuma das pessoas acerta novamente todas as cartas.

**CONCLUSÃO:** se você contar para alguém que esta pessoa tem poderes extra-sensoriais, a pessoa perde seus poderes.

O QUE A TEORIA DAS PROBABILIDADES PODE NOS DIZER?

$P_1(V/C) = 0,5$  (1/2 cores pode sair)

$P_1 = 0,5 = P_2 = P_3 = \dots = P_i$

CÁLCULO:  $P(P_1 \text{ e } P_2) = 0,5 * 0,5$

$P(P_1 \text{ e } P_2 \text{ e } \dots \text{ e } P_{10}) = (0,5)^{10} = 1/(2^{10}) = 0,000977$   
 $= 0,0977\% \sim 0,1\%$

## PROBABILIDADE COMBINADA

Podemos analisar a probabilidade combinada de um evento com duas características com auxílio de uma tabela.

Obs: chamamos de Tabela de Contingência.

		Característica 2		
Característica 1		B	B'	Sub Total
	A	$P(A \text{ e } B)$	$P(A \text{ e } B')$	$P(A)$
	A'	$P(A' \text{ e } B)$	$P(A' \text{ e } B')$	$P(A')$
	Sub Total	$P(B)$	$P(B')$	$\square P = 1$

## EXEMPLO

Suponha uma pesquisa feita em uma loja, onde se pergunta:  
(1) você deseja comprar uma TV? E ao passar no caixa,  
se efetivamente comprou uma TV.

$$P(A \text{ e } B) = ?$$

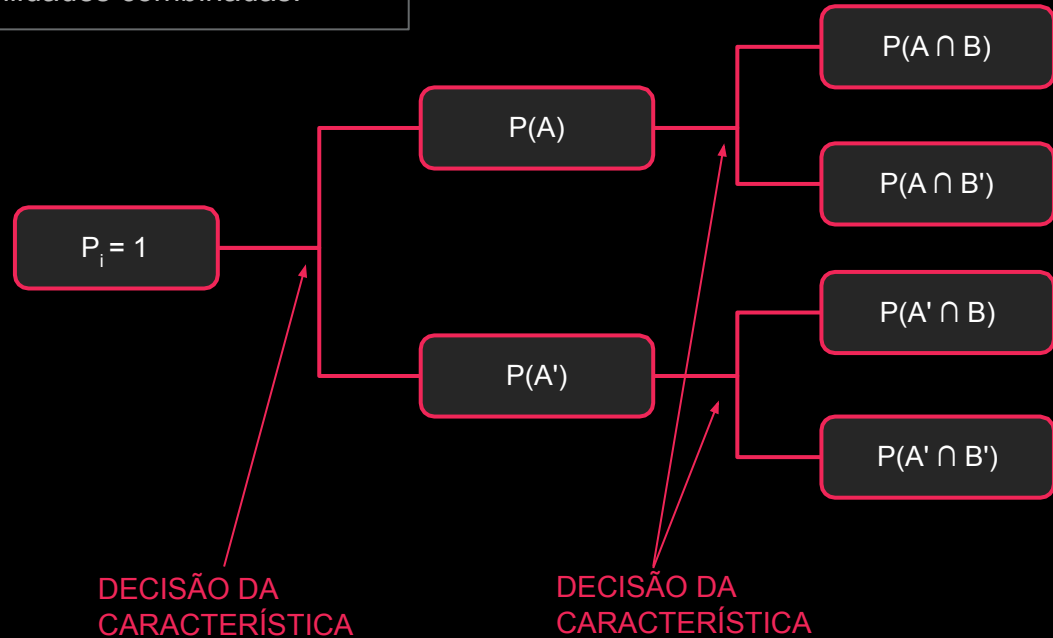
$$P(A \text{ ou } B) = ?$$

Independentes estatisticamente?

	Característica 2			
Característica 1		Sim (B)	Não (B')	Sub Total
	Sim (A)	200/1000	50/1000	250/1000
	Não (A')	100/1000	650/1000	750/1000
	Sub Total	300/1000	700/1000	1000

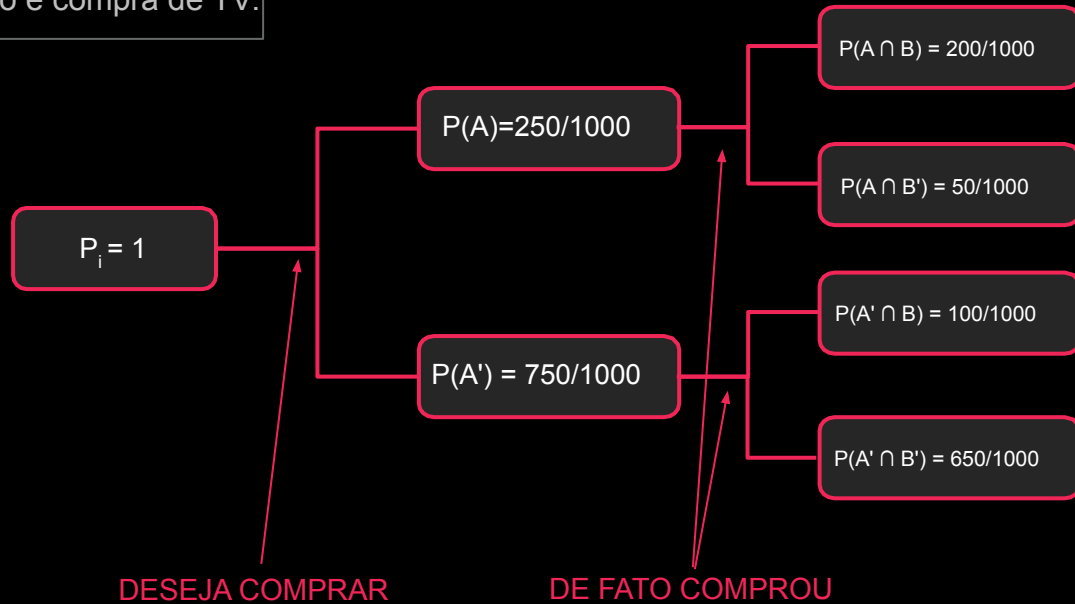
## ÁRVORE DE DECISÃO

É uma representação de probabilidades combinadas:



## EXEMPLO

Para o exemplo do desejo e compra de TV:



## EXEMPLO

Preencha a tabela de contingência para o exercício do Titanic.

$P(A \text{ e } B) = ?$

$P(A \text{ ou } B) = ?$

Independentes estatisticamente?

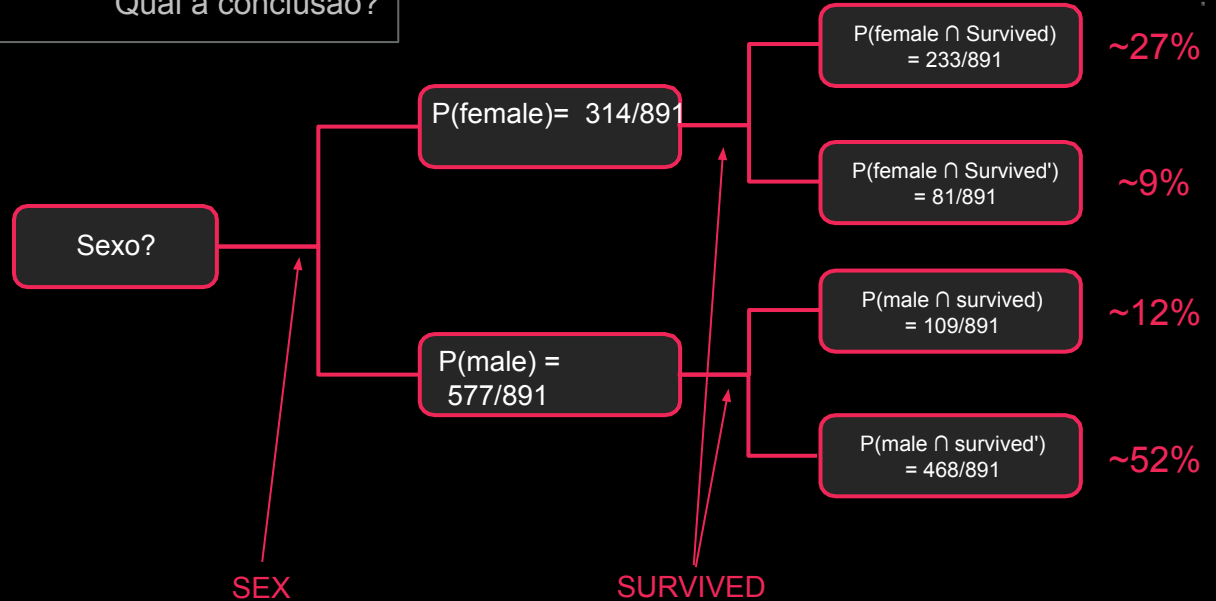
dica: `count(d,c("Sex", "Survived"))`

	SURVIVED ?			
SEX		Sim, 1 (B)	Não (B')	Sub Total
	Female (A)	233/891	81/891	314/891
	Male (A')	109/891	468/891	577/891
	Sub Total	342/891	549/891	891/891

## ÁRVORE DE DECISÃO - EXERCÍCIO

Para o exemplo do Titanic, sobreviventes, Sexo:

Qual a conclusão?



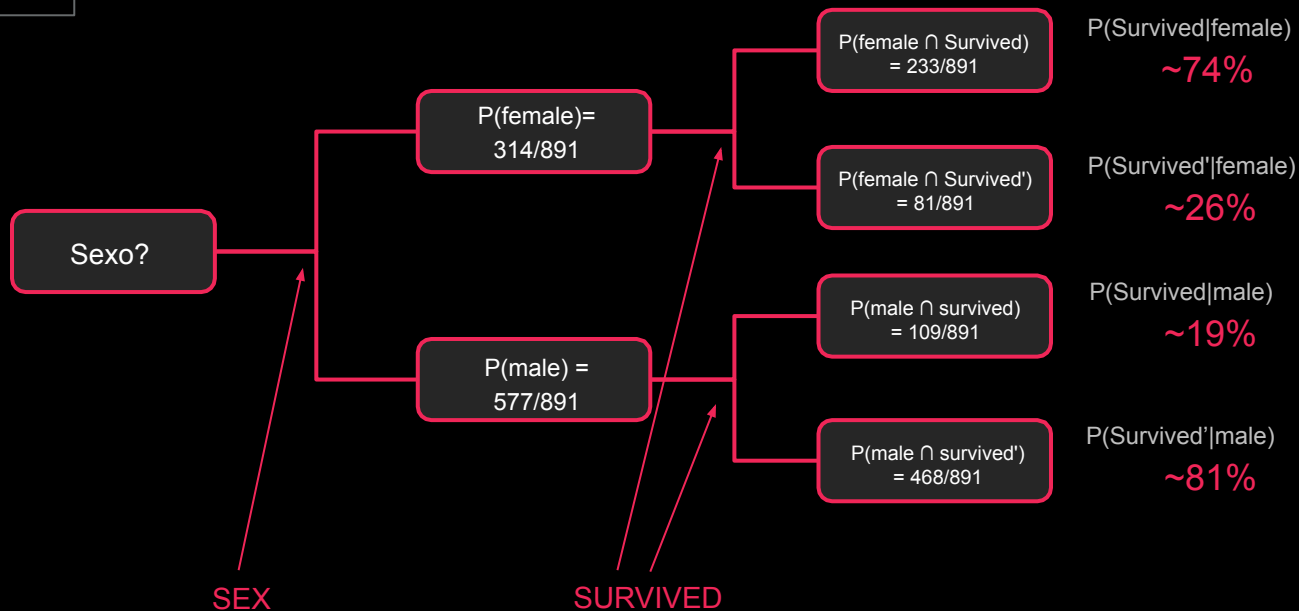


## ÁRVORE DE DECISÃO - EXERCÍCIO

Qual a probabilidade de sobrevivência, dado que (condição) o passageiro seja do sexo feminino?

Probabilidade condicional!  $P(A|B) = ?$   $P(\text{Survived}|\text{female}) = ?$

Qual a conclusão?



ÁRVORE DE DECISÃO:

VAMOS  
TENTAR?

Como testar uma variável?

```
ifelse(mtcars$mpg>20,"Economico","Gastao")
```

Como unir uma coluna a um dataset?

```
m1<-cbind(mtcars,ifelse(mtcars$mpg>20,"Eco","Gastao"))
```

Gerando os resultados para Titanic: só mulheres sobrevivem!

```
dsPREDICT <- read.csv(file.choose())
```

```
p<-dsPREDICT
```

```
p<-cbind(p$PassengerId, ifelse(p$Sex=="female",1,0)) _
```

```
colnames(p) <- c("PassengerId", "Survived") _
```

```
write.csv(p, "predict.csv", row.names=FALSE) _
```

ÁRVORE DE DECISÃO:

## COMO ENCONTRAR?

Fizemos um exercício com uma variável;  
quantas possibilidades existem?  
E se combinarmos as variáveis?  
=> Algoritmos !!!

```
install.packages("party")  
library(party)  
mcDT <- ctree(Survived ~ Sex, data=dsTRAIN)  
plot(mcDT, type="simple")
```



# AVALIANDO E COMPARANDO MODELOS DE CLASSIFICAÇÃO

## AVALIAÇÃO DE MODELOS

Como avaliar se um modelo prevê corretamente os dados?

	PREVISTO		
REAL		PVC	NORMAL
	PVC	0.78	0.22
	NORMAL	0.16	0.84

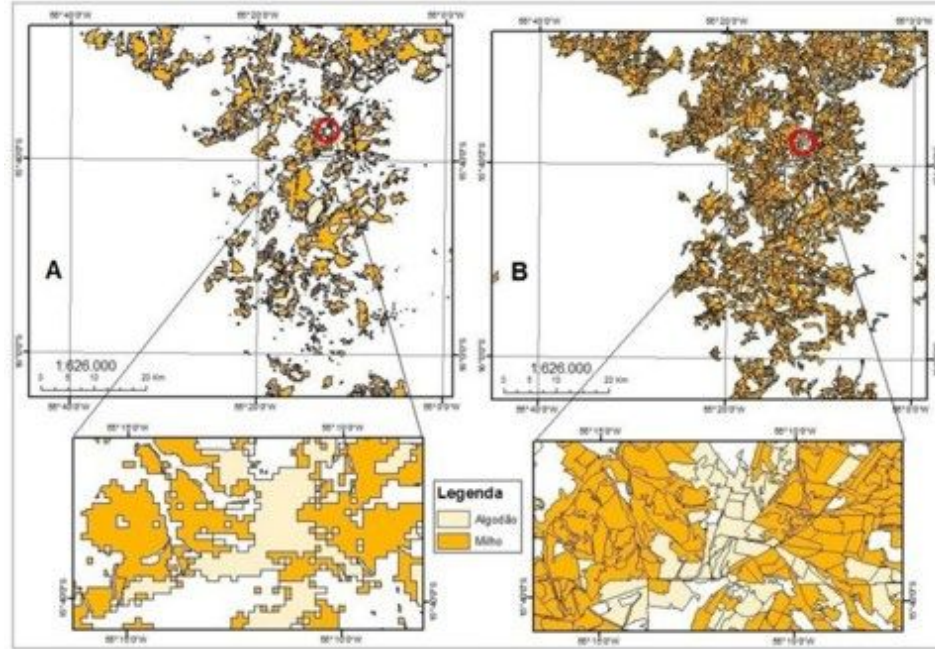
MATRIZ DE CONFUSÃO DOS RESULTADOS

## AVALIAÇÃO DE MODELOS

Obs.1: na fonte não é citado se o real é coluna ou linha.  
Inseri arbitrariamente para utilizar o exemplo.

Obs.2: reparem que os dados foram apresentados como percentual (probabilidade) em linha. Isto é, 78% dos PVCs reais foram previsto corretamente. 22% foram estimados erroneamente como normais.

Fonte: <http://slideplayer.com.br/slide/1266306/>  
Classificação de arritmias cardíacas



Fonte: MAPEAMENTO DE CULTURAS AGRÍCOLAS POR IMAGENS DE SATÉLITE - COSTA, SOUTO e Zeilhofer

<http://www.sinageo.org.br/2012/trabalhos/10/10-379-577.html>

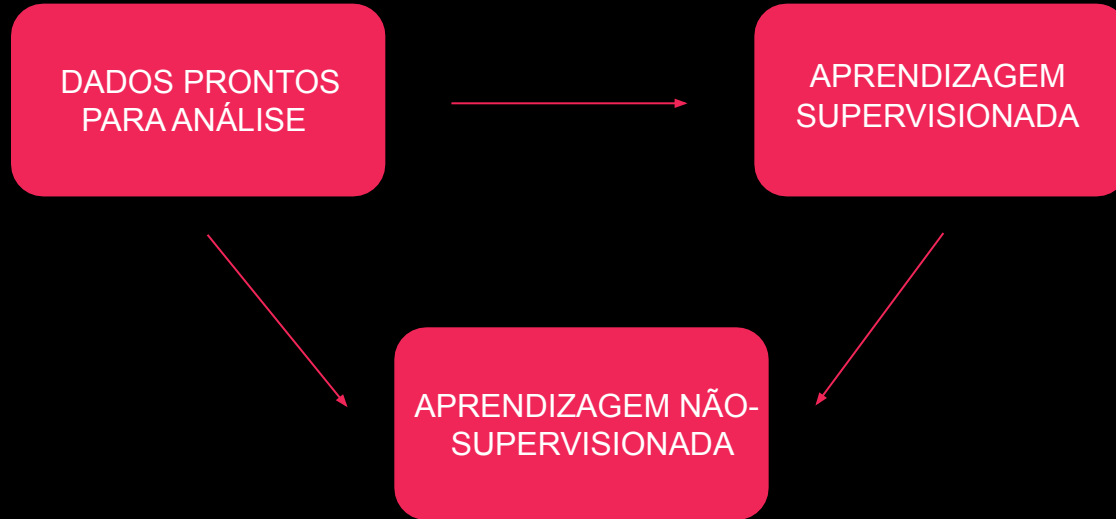
	VERDADE DE CAMPO				
CLASSIFICADOR		ALGODÃO	MILHO	OUTROS	TOTAL
	ALGODÃO	5.569	819	0	6.388
	MILHO	0	3.552	0	3.552
	OUTROS	0	761	5.674	6.435
	TOTAL	5.569	5.119	5.674	16.375



# CONCEITO DE APRENDIZAGEM

- APRENDIZAGEM SUPERVISIONADA
  - Regressão
  - Classificação
- APRENDIZAGEM NÃO-SUPERVISIONADA
  - Clusterização
  - Redução de dimensão
  - Descoberta de padrões
  - ETC

# SUPERVISIONADA x NÃO-SUPERVISIONADA



# DATA MINING

“Descobrir, gerar conhecimento baseado em dados. Propicia análise automatizada e escalável para grandes conjuntos de dados.”

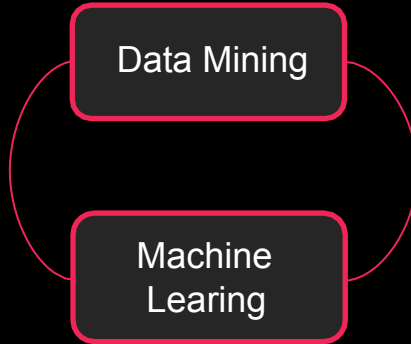
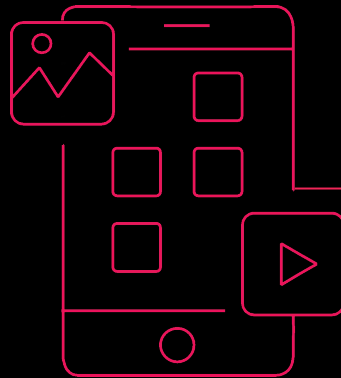
Prof. Jiawei Han,  
UI

# MACHINE LEARNING

“Estudo e construção de métodos para aprender sobre dados e fazer previsões.”

Prof. Ameet Talwalkar,  
UCLA

# MACHINE LEARNING X DATA MINING



ESTUDOS PONTUAIS

CONHECIMENTO

METODOLOGIAS

APLICAÇÕES

EM GERAL...

# Obrigado



[profDiogenes.Justo@fiap.com.br](mailto:profDiogenes.Justo@fiap.com.br)

SHIFT  
FIAP

Copyright © 2018 | Diógenes Justo

Todos os direitos reservados. Reprodução ou divulgação total ou parcial deste documento, é expressamente proibido sem consentimento formal, por escrito, do professor/autor.