

# SPORTS ANALYTICS

Mudando o Jogo

Ígor Barbosa da Costa  
Carlos Eduardo Pires  
Leandro Balby Marinho



# Apresentação

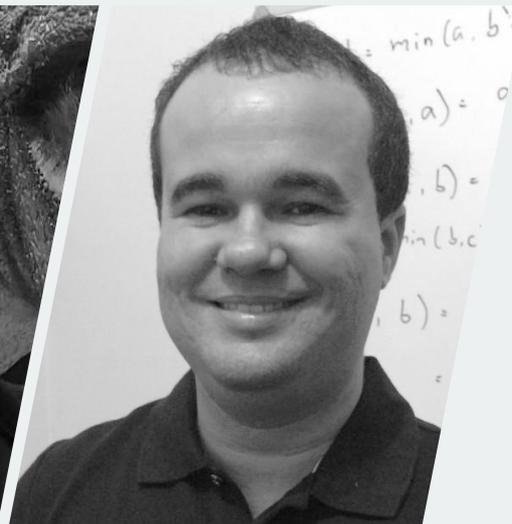
## Quem são os autores?



**IGOR BARBOSA  
DA COSTA**



**CARLOS  
EDUARDO PIRES**



**LEANDRO BALBY  
MARINHO**

# Apresentação

## Os Autores

**Ígor Barbosa da Costa** é professor de computação do Instituto Federal de Educação, Ciência e Tecnologia da Paraíba (IFPB), campus Campina Grande. Graduado em Ciência da Computação pela Universidade Federal de Campina Grande - UFCG (2006) e Mestre pela Universidade Federal de Pernambuco - UFPE (2010). Tem experiência na área de Ciência da Computação, com ênfase em Bancos de Dados e Desenvolvimento, atuando principalmente nos seguintes temas: Mineração de Dados e Descoberta de Conhecimento (com foco atual em dados esportivos).



# Apresentação

## Os Autores

**Carlos Eduardo Santos Pires** concluiu a Graduação em Ciência da Computação, em 1997, pela Universidade Federal de Campina Grande (UFCG) e Mestrado em Informática, em 2000, pela mesma instituição. Em 2009, concluiu o Doutorado na Universidade Federal de Pernambuco (UFPE), tendo realizado Doutorado-Sanduiche na Université de Versailles, na França. Atualmente é Professor Adjunto do Departamento de Sistemas e Computação (DSC), da UFCG. Tem experiência na área de Ciência da Computação, com ênfase em Bancos de Dados, atuando nos seguintes temas: Qualidade de Dados, Integração de Dados, Descoberta de Conhecimento e Big Data.



# Apresentação

## Os Autores

**Leandro Balby Marinho** é Doutor em Ciência da Computação, pela Universidade de Hildesheim, Alemanha, 2010. Mestre em Engenharia Elétrica, UFMA, Brasil, 2005. Bacharel em Ciência da Computação, UFMA, Brasil, 2002. Professor do Departamento de Sistemas e Computação da Universidade Federal de Campina Grande (UFCG). Atua como docente, pesquisador e orientador nos cursos de graduação e pós-graduação em ciência da computação. Suas áreas de especialização são: Inteligência Artificial, Mineração de Dados e Recuperação da Informação.



# Apresentação

## Os Autores



igor.costa@ifpb.edu.br  
cesp@dsc.ufcg.edu.br  
lbmarinho@dsc.ufcg.edu.br

# Apresentação

## A Metodologia



ESPORTE



MINERAÇÃO DE  
DADOS



TEMA DE  
DOUTORADO



NOVOS  
PESQUISADORES

# Apresentação

## A Metodologia



EXPOSITIVA



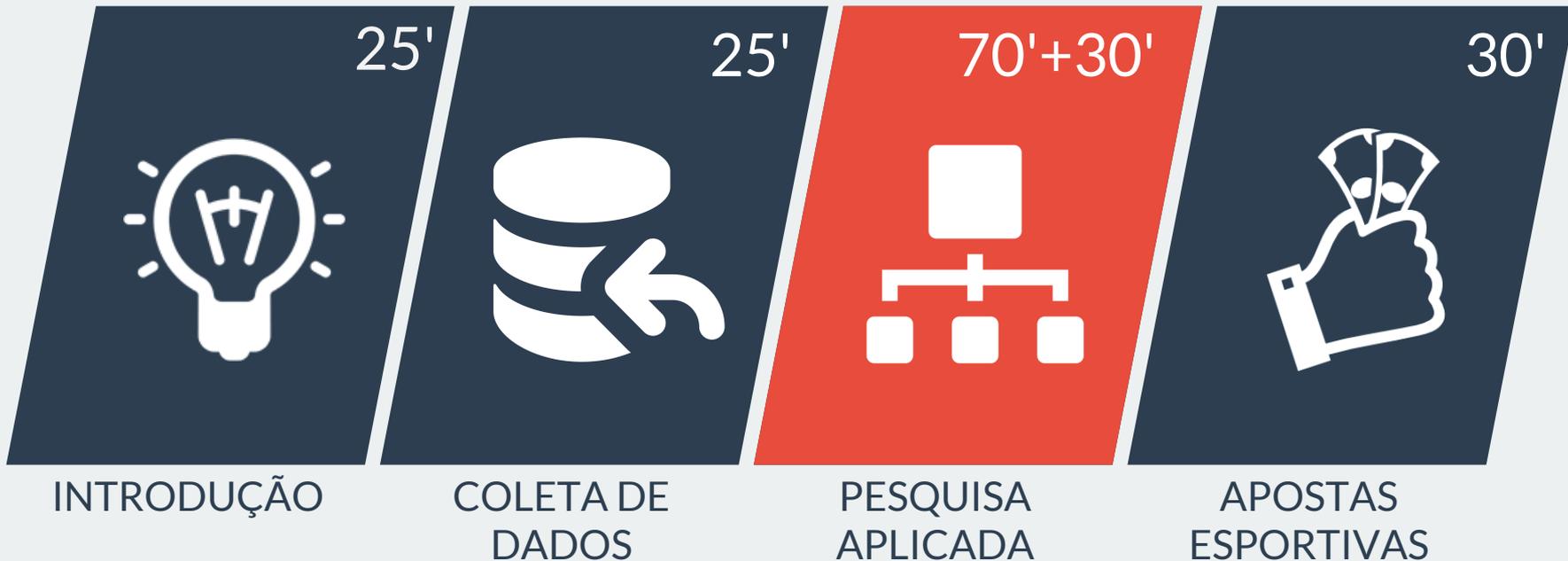
COLABORATIVA



PRÁTICA

# Apresentação

## Agenda do Minicurso



# Apresentação

## O Público



CIENTISTA  
DE DADOS



FÃ DE  
ESPORTES



APOSTADOR



CURIOSO





# 01

# Introdução

*“Until recently, it was very much about collecting data on **what** had happened, without looking at **why** it had happened.”*

**PAUL POWER**  
Cientista de Dados da Prozone/STATS



# Motivação Histórica

## Século XX: Regulamentação e Profissionalização



**FIBA**

We Are Basketball

BASQUETE



**FIFA**

FUTEBOL



**FIVB**

VÔLEI



**WORLD  
RUGBY™**

RUGBY

# Motivação Histórica

## Crescimento da Imprensa Esportiva

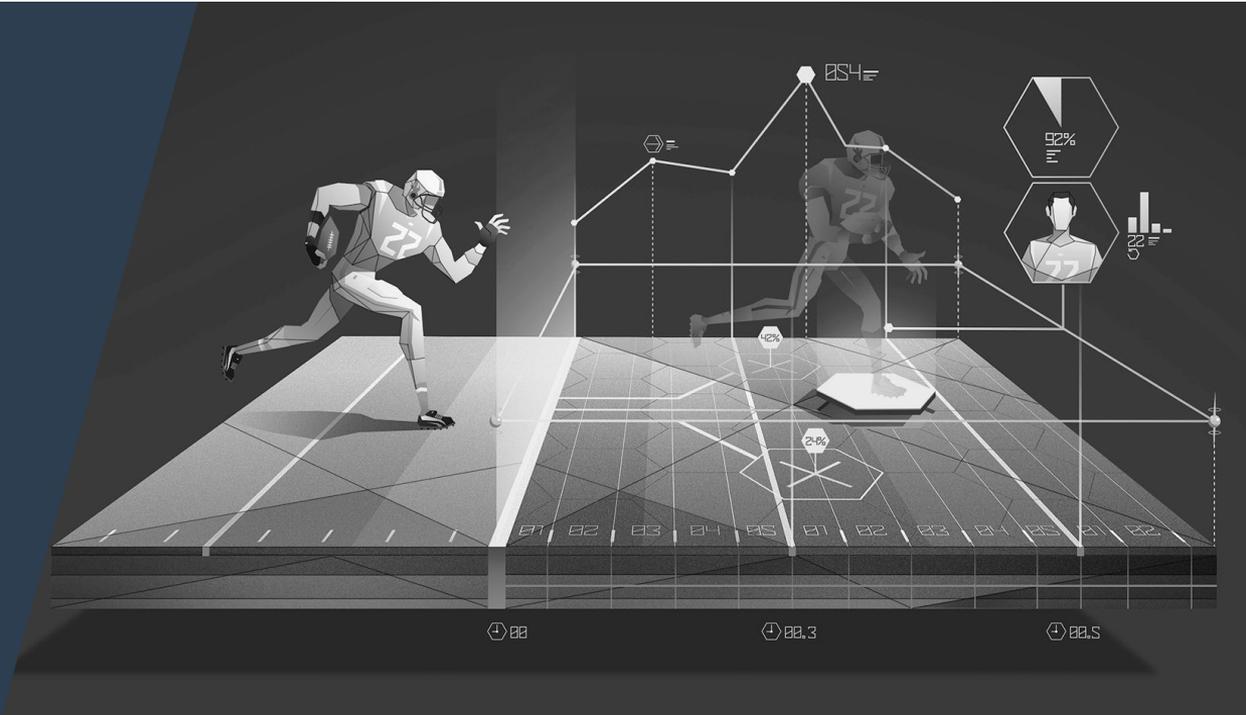


# Motivação Histórica

## Sports Analytics como Vantagem Competitiva



# SPORTS ANALYTICS



# Motivação Histórica

## Os Pioneiros



CHARLES REEP



BILLY JAMES

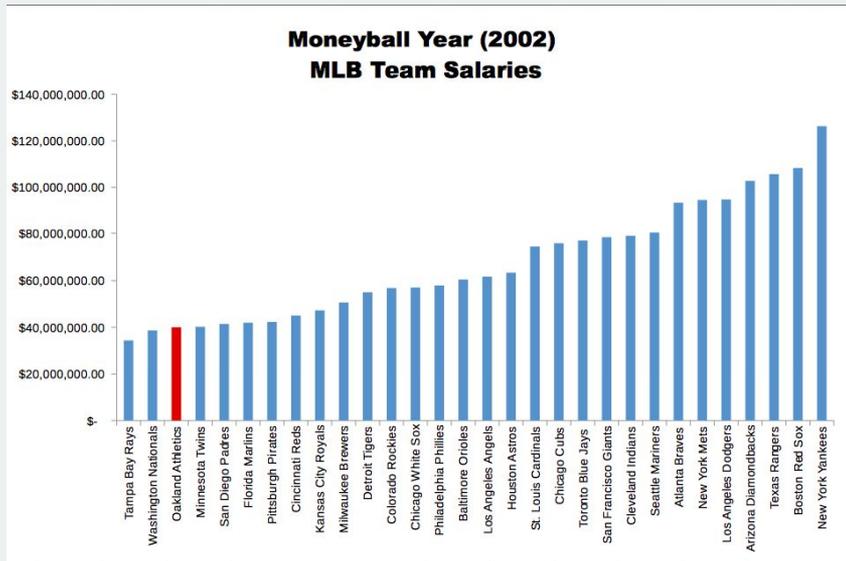


DEAN OLIVER



# Motivação Histórica

## O Caso de Sucesso do Oakland Athletics



# Motivação Histórica

## O Caso de Sucesso do Oakland Athletics



MLB.com

2002

20 vitórias consecutivas  
Recorde na Liga Americana da  
MLB



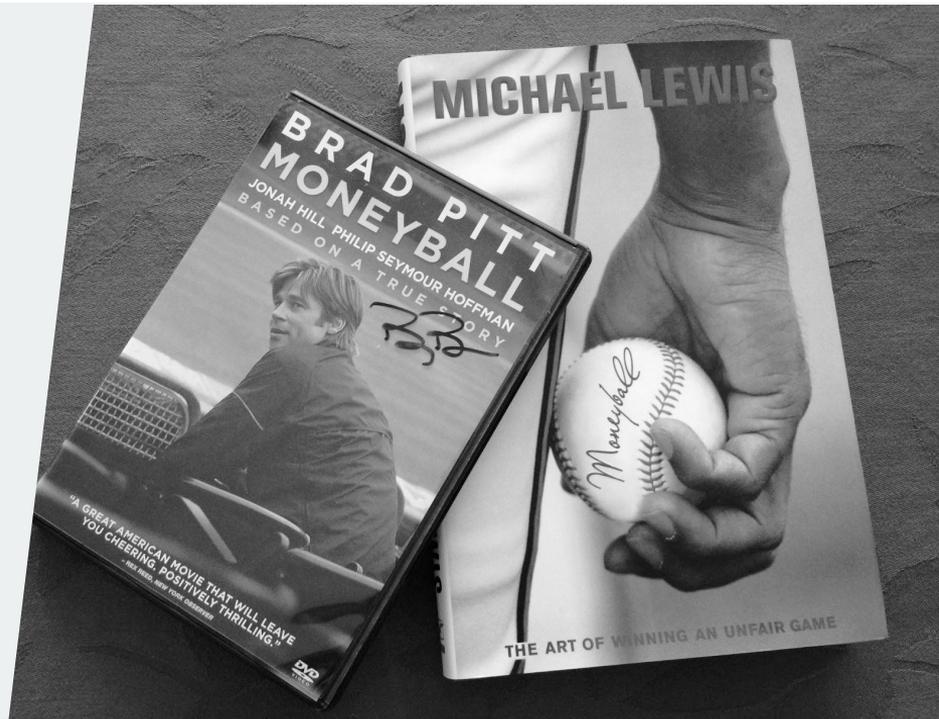
# Motivação Histórica

## O Impacto de Moneyball



### Moneyball

The Art of Winning an Unfair Game



# Sports Analytics: Taxonomy 1.0

## Classificação dos Objetivos de Pesquisa



COMPETIÇÃO



LAZER



EVENTO  
INCERTO

“Analytics don’t work at all. It’s just some crap that people who were really smart made up to try to get in the game because they had no talent. (...) so they made up a term called *analytics*.”

**CHARLES BARKLEY**

Ex-jogador e comentarista da TNT

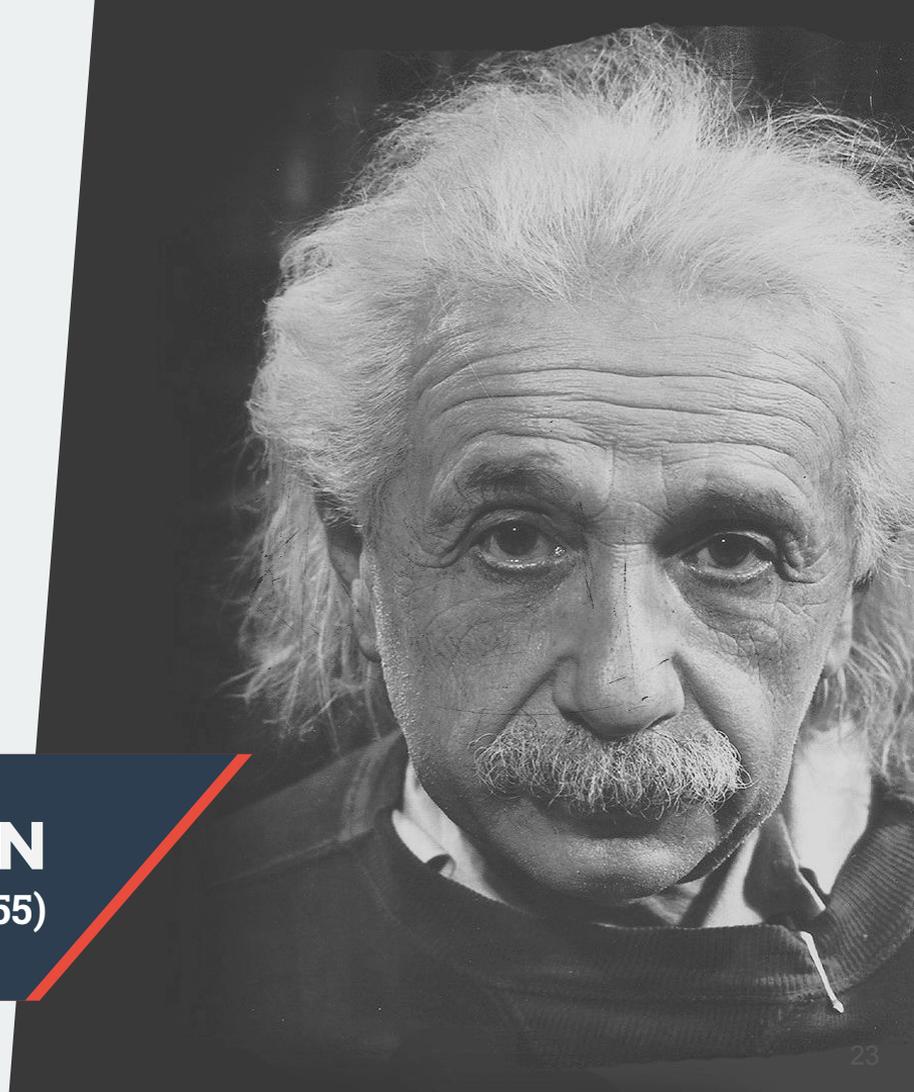




“Not everything that can be counted *counts*, and not everything that counts *can be counted*.”

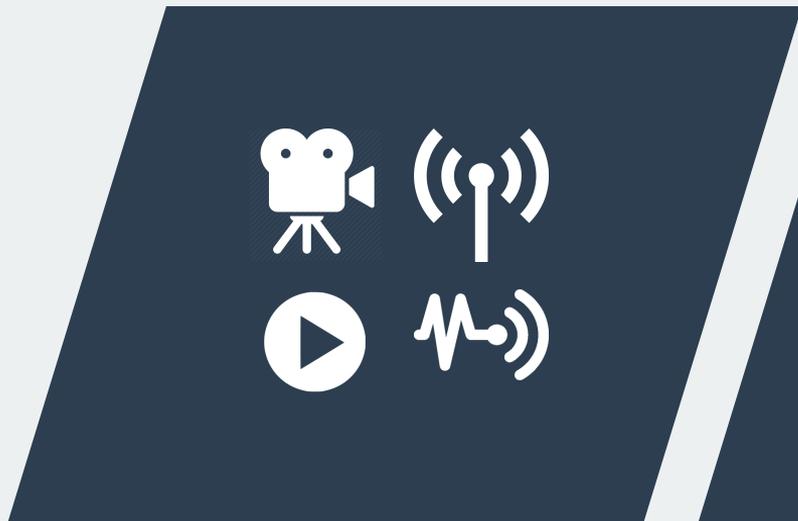
**ALBERT EINSTEIN**

Físico teórico (1879-1955)



# Coleta de Dados

## A Origem dos Dados



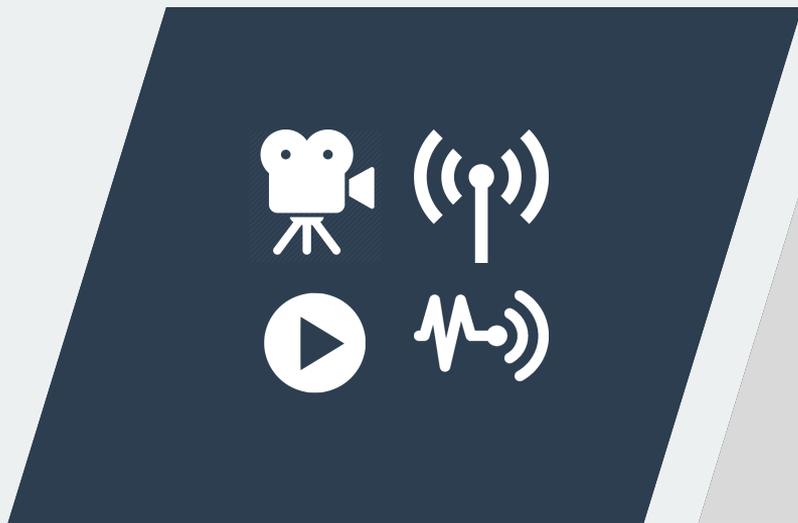
VÍDEOS E SENSORES



CONTEXTO

# Coleta de Dados

## A Origem dos Dados



VÍDEOS E SENSORES



CONTEXTO

# Dados de Vídeos e Sensores

## Dispositivos de Coleta



SISTEMAS MULTI-CÂMERAS



WEREABLES

# Dados de Vídeos e Sensores

## Formas de Representação



DADOS DE  
MOVIMENTAÇÃO



DADOS DE  
EVENTOS



DADOS  
DESCRITIVOS

# Dados de Vídeos e Sensores

## Formas de Representação



DADOS DE  
MOVIMENTAÇÃO



DADOS DE  
EVENTOS

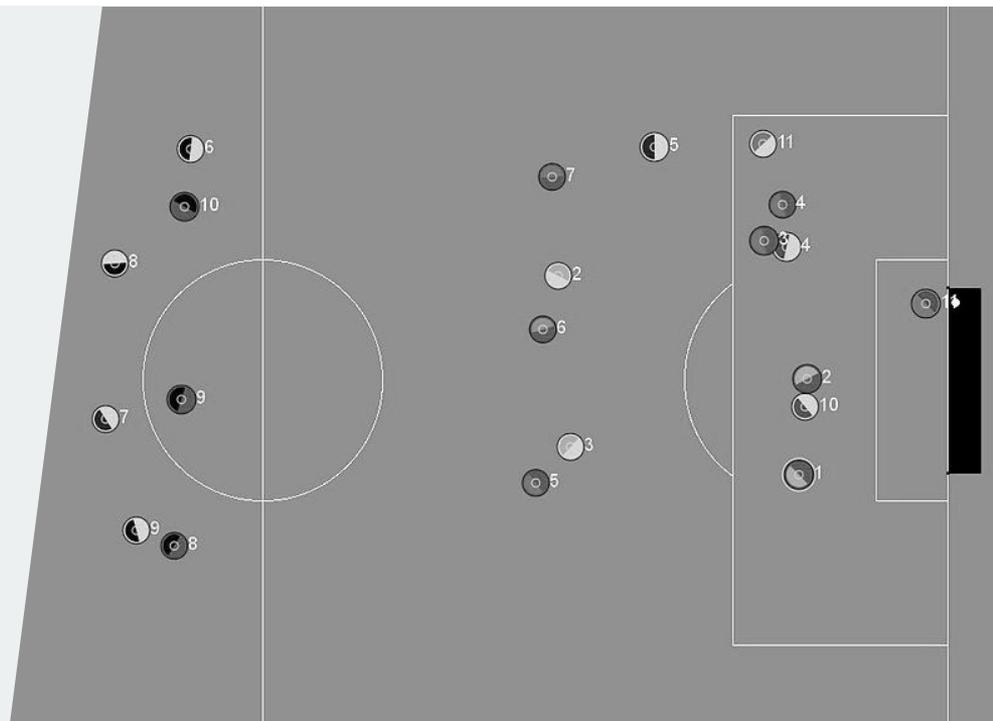


DADOS  
DESCRITIVOS

# Dados de Vídeos e Sensores

## Dados de Movimentação

Dados de movimentação ou *espaço-temporais* descrevem onde um jogador ou objeto está localizado em um momento específico.



# Dados de Vídeos e Sensores

## Dados de Movimentação

# STATS

*“The STATS SportVU tracking system delivers performance statistics by extracting and processing coordinates of players (X,Y) and the ball (X,Y,Z) through HD cameras as well as sophisticated software and statistical algorithms.”*



# Dados de Vídeos e Sensores

## Formas de Representação



DADOS DE  
MOVIMENTAÇÃO



DADOS DE  
EVENTOS



DADOS  
DESCRITIVOS

# Dados de Vídeos e Sensores

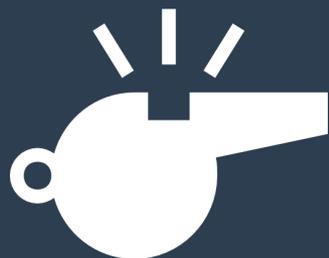
## Dados de Eventos

*Evento é um acontecimento relevante em uma disputa esportiva. Uma disputa pode ser descrita como uma sequência ordenada de eventos.*



# Dados de Vídeos e Sensores

## Dados de Eventos



BASEADO NAS REGRAS



BASEADO NA INTERAÇÃO

# Dados de Vídeos e Sensores

## Dados de Eventos



# Dados de Vídeos e Sensores

## Formas de Representação



DADOS DE  
MOVIMENTAÇÃO



DADOS DE  
EVENTOS



DADOS  
DESCRITIVOS

# Dados de Vídeos e Sensores

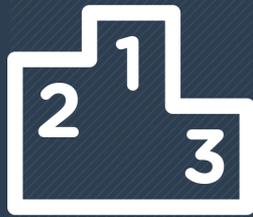
## Dados Descritivos

Os dados *descritivos* incluem tudo que pode ser contado ou medido durante as competições.

| CLASSIFICAÇÃO |             | P | J  | V  | E  | D | GP | GC | SG | %   | ÚLT. JOGOS |       |
|---------------|-------------|---|----|----|----|---|----|----|----|-----|------------|-------|
| 1             | Corinthians | 0 | 53 | 24 | 16 | 5 | 3  | 34 | 13 | 21  | 73.6       | ●●●●● |
| 2             | Grêmio      | 0 | 43 | 24 | 13 | 4 | 7  | 40 | 21 | 19  | 59.7       | ●●●●● |
| 3             | Santos      | 0 | 41 | 24 | 11 | 8 | 5  | 25 | 16 | 9   | 56.9       | ●●●●● |
| 4             | Palmeiras   | 0 | 40 | 24 | 12 | 4 | 8  | 35 | 26 | 9   | 55.6       | ●●●●● |
| 5             | Flamengo    | 0 | 38 | 24 | 10 | 8 | 6  | 33 | 23 | 10  | 52.8       | ●●●●● |
| 6             | Cruzeiro    | 0 | 37 | 24 | 10 | 7 | 7  | 29 | 21 | 8   | 51.4       | ●●●●● |
| 7             | Botafogo    | 0 | 37 | 24 | 10 | 7 | 7  | 29 | 23 | 6   | 51.4       | ●●●●● |
| 8             | Atlético-PR | 1 | 34 | 24 | 9  | 7 | 8  | 29 | 27 | 2   | 47.2       | ●●●●● |
| 9             | Vasco       | 1 | 31 | 24 | 9  | 4 | 11 | 24 | 35 | -11 | 43.1       | ●●●●● |
| 10            | Atlético-MG | 1 | 31 | 24 | 8  | 7 | 9  | 26 | 28 | -2  | 43.1       | ●●●●● |

# Dados de Vídeos e Sensores

## Dados Descritivos



RANKINGS



SCOUTS



ATRIBUTOS DOS  
JOGADORES



OUTROS

# Coleta de Dados

## A Origem dos Dados



VÍDEOS E SENSORES



CONTEXTO

# Dados de Contexto

## Conceito

*Dados adicionais que relacionados aos eventos esportivos “indiretamente”.*



# Dados de Contexto

## Exemplos



AMBIENTE



OPINIÕES



APOSTAS



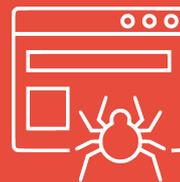
OUTROS

# Coleta de Dados

## Como ter acesso aos Dados?



DADOS ESTRUTURADOS



CRAWLERS E  
SCRAPERS

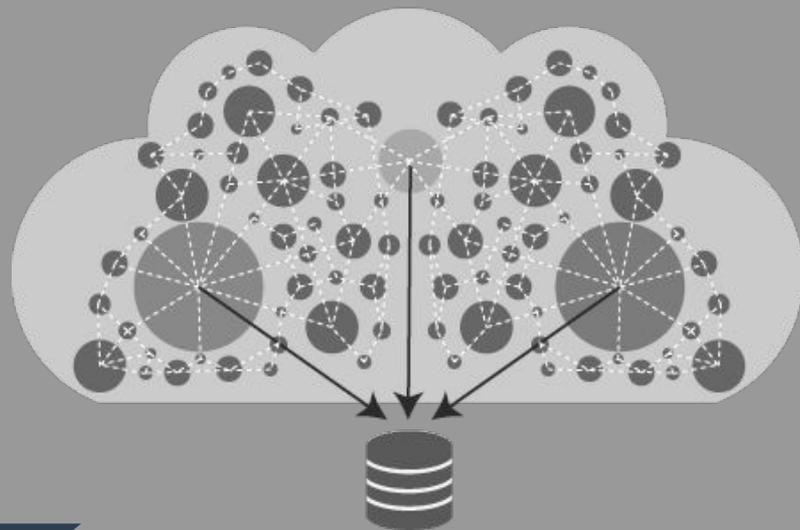


# Acesso aos Dados

## Web Crawler (Rastreamento)

*Rastreamento é a tarefa de navegar pela Internet de forma automatizada para indexar páginas nas quais os dados desejados estão embutidos.*

web crawler



# Acesso aos Dados

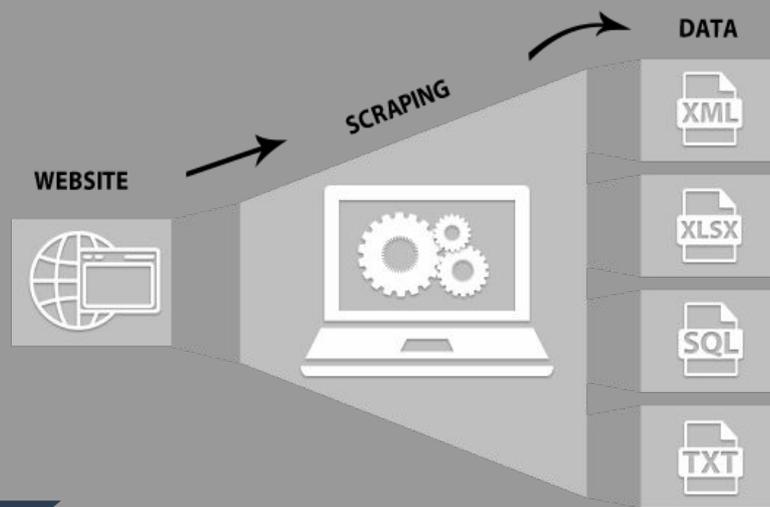
## Web Crawler (Rastreamento)

```
import requests
r = requests.get('http://www.espn.com/mlb/statistics')
print(r.status_code)
#200
print(r.text)
#...
#<div class="span-6">
#   <div class="mod-container mod-no-header-footer mod-page-#header">
#     <div class="mod-content">
#       <h1 class="h2">MLB Statistics - 2017</h1>
#     <div class="floatleft">
#...
```

# Acesso aos Dados

## Web Scraper (Raspagem)

*Raspagem é a tarefa de extrair informações específicas de documentos web.*



*web scraper*

# Acesso aos Dados

## Web Scraper (Raspagem)

```
from bs4 import BeautifulSoup
soup = BeautifulSoup(html_doc, 'html.parser')

print(soup.prettify())
# <html>
# <head>
# <title>
#   Exemplo
# </title>
# </head>
# <body>
# <p class="estilo">
#   <b>
#     Pagina de Exemplo
#   </b>
# </p>
# </p>
# </body>
# </html>
```

# Acesso aos Dados

## Web Scraper (Raspagem)

```
print(soup.title)
# <title>Exemplo</title>

print(soup.title.name)
# u'Exemplo'

print(soup.p)
# <p class="title"><b>Pagina de Exemplo</b></p>

print(soup.p['class'])
# u'Estilo'
```

# 03

## Pesquisa Aplicada

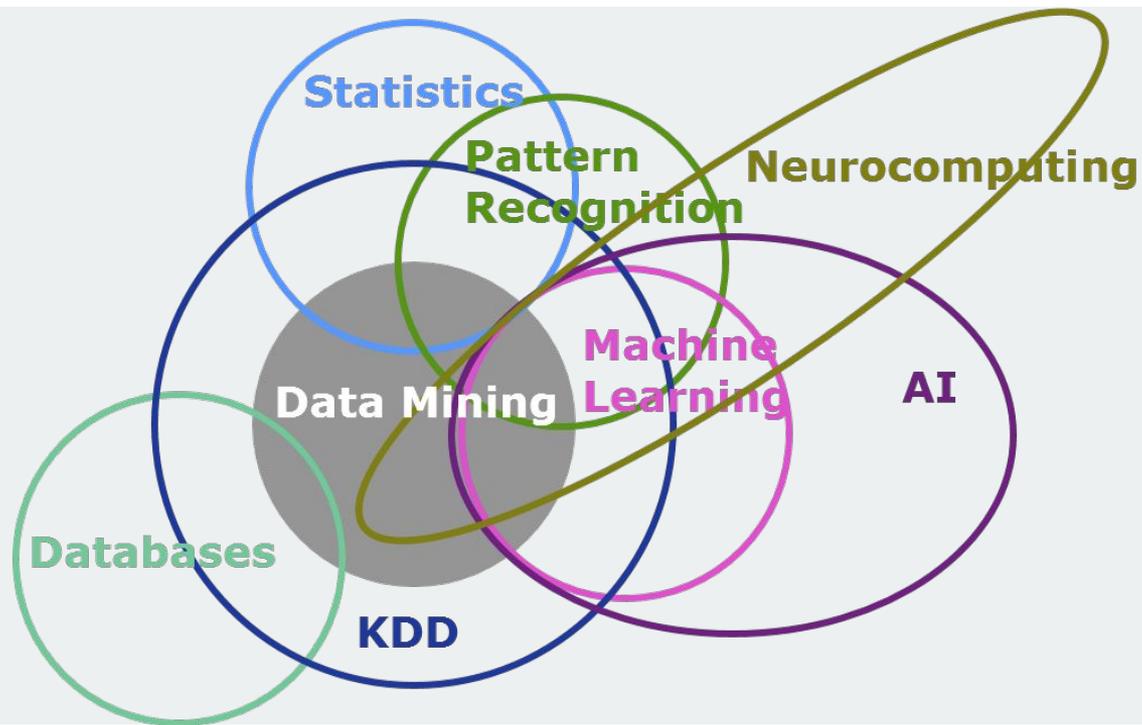
WE WATCH.

WE ASK.



# Pesquisa Aplicada

## Contextualização



# Pesquisa Aplicada

## Processo KDD

### Knowledge Discovery in Databases



# Pesquisa Aplicada

## Processo KDD

### Knowledge Discovery in Databases



# Seleção de Dados

## Conceitos



### VARIÁVEIS

*atributos, fatores ou características (features)*



### REGISTROS

*casos, objetos ou observações*

# Seleção de Dados

## Conceitos

| <i>variáveis</i> | NROD | TIMEA       | TIMEB       | GA | GB | LOCAL             |
|------------------|------|-------------|-------------|----|----|-------------------|
|                  | 24   | Botafogo    | Santos      | 2  | 0  | Rio de Janeiro-RJ |
|                  | 24   | Avai        | Atlético-MG | 1  | 1  | Florianópolis-SC  |
| <i>registros</i> | 24   | Flamengo    | Sport       | 2  | 0  | Rio de Janeiro-RJ |
|                  | 24   | Corinthians | Vasco       | 1  | 0  | São Paulo-SP      |

# Seleção de Dados

## Exemplo

*Gostaríamos de realizar  
previsões nos jogos dos  
campeonato brasileiro.*

*O que deveríamos selecionar?  
O que está disponível?*



# Seleção de Dados

## Exemplo

*Número da Rodada*  
*Nome do Time A*  
*Nome do Time B*  
*Gols de A*  
*Gols de B*  
*Local da Partida*



# Seleção de Dados

## Exemplo

| NROD | TIMEA       | TIMEB       | GA  | GB  | LOCAL             |
|------|-------------|-------------|-----|-----|-------------------|
| ...  | ...         | ...         | ... | ... | ...               |
| 24   | Botafogo    | Santos      | 2   | 0   | Rio de Janeiro-RJ |
| 24   | Avaí        | Atlético-MG | 1   | 1   | Florianópolis-SC  |
| 24   | Flamengo    | Sport       | 2   | 0   | Rio de Janeiro-RJ |
| 24   | Corinthians | Vasco       | 1   | 0   | São Paulo-SP      |
| ...  | ...         | ...         | ... | ... | ...               |

# Pesquisa Aplicada

## Processo KDD

### Knowledge Discovery in Databases



# Pré-Processamento e Limpeza de Dados

## Etapas



LIMPEZA DE DADOS



PRÉ-PROCESSAMENTO



# Pré-Processamento e Limpeza de Dados

## Etapas



LIMPEZA DE DADOS



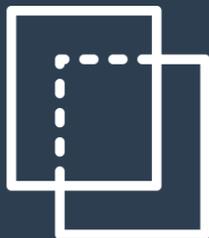
PRÉ-PROCESSAMENTO

# Limpeza de Dados

## Tarefas



AUSENTES



DUPLICADOS



INCONSISTENTES



OUTLIERS

# Limpeza De Dados

## Dados Ausentes

| NROD | TIMEA       | TIMEB       | GA | GB | LOCAL             |
|------|-------------|-------------|----|----|-------------------|
| 24   | Botafogo    | Santos      | 2  | 0  | Rio de Janeiro-RJ |
| 24   | Avaí        | Atlético-MG | 1  | 1  | Florianópolis-SC  |
| 24   | Flamengo    | Sport       | 2  | 0  | Rio de Janeiro-RJ |
| 24   | Corinthians | Vasco       |    | 0  | São Paulo-SP      |
| 24   | Grêmio      | Chapecoense | 0  | 1  | Porto Alegre-RS   |
| 24   | Cruzeiro    | Bahia       | 1  | 0  |                   |

# Limpeza de Dados

## Dados Duplicados

| NROD | TIMEA       | TIMEB       | GA | GB | LOCAL             |
|------|-------------|-------------|----|----|-------------------|
| 24   | Botafogo    | Santos      | 2  | 0  | Rio de Janeiro-RJ |
| 24   | Avaí        | Atlético-MG | 1  | 1  | Florianópolis-SC  |
| 24   | Avaí        | Atlético-MG | 1  | 1  | Florianópolis-SC  |
| 24   | Flamengo    | Sport       | 2  | 0  | Rio de Janeiro-RJ |
| 24   | Corinthians | Vasco       | 1  | 0  | São Paulo-SP      |
| 24   | Grêmio      | Chapecoense | 0  | 1  | Porto Alegre-RS   |
| 24   | Cruzeiro    | Bahia       | 1  | 0  | Belo Horizonte-MG |

# Limpeza de Dados

## Dados Inconsistentes

| NROD | TIMEA       | TIMEB       | GA | GB | LOCAL             |
|------|-------------|-------------|----|----|-------------------|
| 24   | Botafogo    | Santos      | 2  | 0  | Rio de Janeiro-RJ |
| 24   | Avaí        | Atlético-MG | W  | ○  | Florianópolis-SC  |
| 24   | Flamengo    | Sport       | 2  | 0  | Rio de Janeiro-RJ |
| 24   | Corinthians | Vasco       | 1  | 0  | São Paulo-SP      |
| 24   | Grêmio      | Chapecoense | 0  | 1  | Porto Alegre-RS   |
| 24   | Cruzeiro    | Bahia       | 1  | 0  | Belo Horizonte-MG |

# Limpeza de Dados

## Outliers

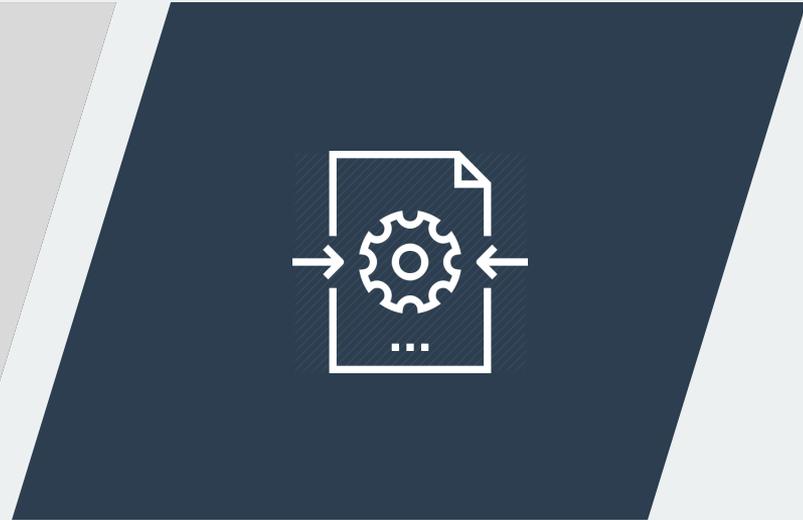
| NROD | TIMEA       | TIMEB       | GA | GB | LOCAL             |
|------|-------------|-------------|----|----|-------------------|
| 24   | Botafogo    | Santos      | 2  | 0  | Rio de Janeiro-RJ |
| 24   | Avaí        | Atlético-MG | 1  | 1  | Florianópolis-SC  |
| 24   | Flamengo    | Sport       | 2  | 0  | Rio de Janeiro-RJ |
| 24   | Corinthians | Vasco       | 1  | 0  | São Paulo-SP      |
| 24   | Grêmio      | Chapecoense | 0  | 1  | Porto Alegre-RS   |
| 24   | Cruzeiro    | Bahia       | 5  | 5  | Belo Horizonte-MG |

# Pré-Processamento e Limpeza de Dados

## Etapas



LIMPEZA DE DADOS



PRÉ-PROCESSAMENTO

# Pré-Processamento

## Tarefas



# Pré-Processamento

## Amostragem

*Abordagem comumente usada para selecionar um subconjunto de registros a serem analisados.*



# Pré-Processamento

## Amostragem

*Qual seria uma boa amostra para prever a média de gols do campeonato Brasileiro de 2017?*



# Pré-Processamento

## Amostragem

*Aoki, Assunção & Melo (2017)  
buscaram medir a influência  
da sorte em alguns esportes.*

*Qual seria uma boa amostra?*



# Pré-Processamento

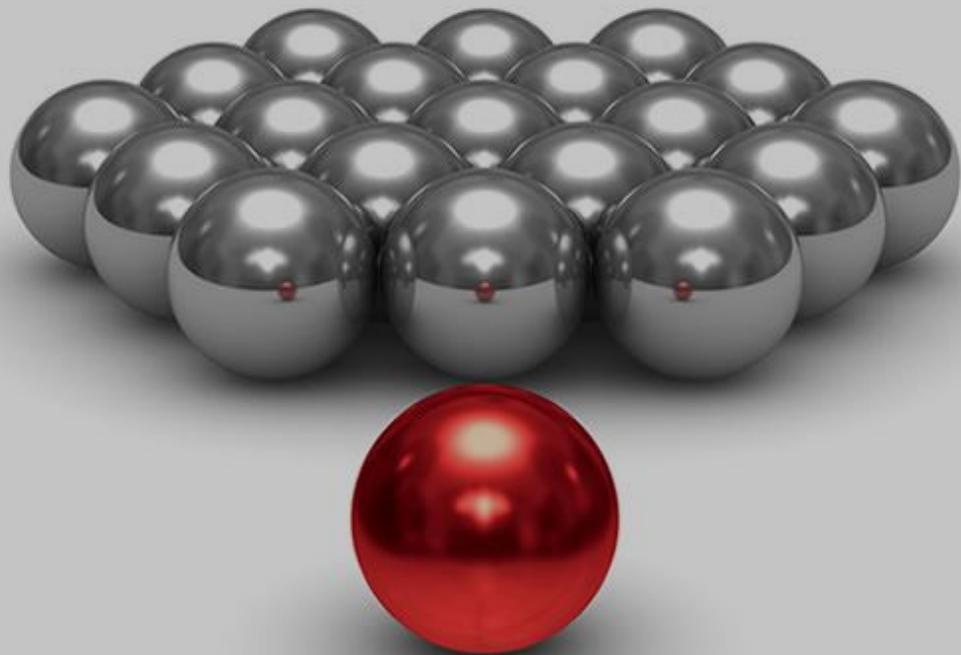
## Amostragem



# Pré-Processamento

## Agregação

*Combinação de vários registros em um único.*



# Pré-Processamento

## Agregação

| NROD | TIMEA       | TIMEB       | GA | GB | LOCAL             |
|------|-------------|-------------|----|----|-------------------|
| 24   | Botafogo    | Santos      | 2  | 0  | Rio de Janeiro-RJ |
| 24   | Avaí        | Atlético-MG | 1  | 1  | Florianópolis-SC  |
| 24   | Flamengo    | Sport       | 2  | 0  | Rio de Janeiro-RJ |
| 24   | Corinthians | Vasco       | 1  | 0  | São Paulo-SP      |



| NROD | NGOALS |
|------|--------|
| 24   | 7      |

# Pré-Processamento

## Criação de Novos Atributos

*Derivar novos atributos a partir dos existentes.*

*Tarefa que depende bastante dos “insights” do pesquisador para resolver o problema proposto.*



# Pré-Processamento

## Criação de Novos Atributos

### *Atributos Atuais*

*Número da Rodada*

*Nome do Time A*

*Nome do Time B*

*Gols de A*

*Gols de B*

*Local da Partida*



# Pré-Processamento

## Criação de Novos Atributos

*Gostaríamos de prever os resultados no Campeonato Brasileiro de 2017.*

*Quais novos atributos poderíamos criar?*



# Pré-Processamento

## Criação de Novos Atributos

| Time A      | Time B      | GA | GB | Local             | %A   | %B   | RES |
|-------------|-------------|----|----|-------------------|------|------|-----|
| Botafogo    | Santos      | 2  | 0  | Rio de Janeiro-RJ | 51.4 | 56.9 | H   |
| Avaí        | Atlético-MG | 1  | 1  | Florianópolis-SC  | 40.3 | 43.1 | D   |
| Flamengo    | Sport       | 2  | 0  | Rio de Janeiro-RJ | 52.8 | 40.3 | H   |
| Corinthians | Vasco       | 1  | 0  | São Paulo-SP      | 73.6 | 43.1 | H   |
| Grêmio      | Chapecoense | 0  | 1  | Porto Alegre-RS   | 59.7 | 38.9 | A   |
| Cruzeiro    | Bahia       | 1  | 0  | Belo Horizonte-MG | 51.4 | 37.5 | H   |

# Pré-Processamento

## Criação de Novos Atributos

### *Atributos Atuais*

*Número da Rodada*

*Nome do Time A*

*Nome do Time B*

*Gols de A*

*Gols de B*

*Local da Partida*



# Pré-Processamento

## Criação de Novos Atributos

*A partir dos resultados coletados, podemos criar uma diversidade de variáveis relacionadas ao desempenho dos clubes que podem ser mais significativas para um modelo de predição.*



# Pré-Processamento

## Criação de Novos Atributos

### *Novos Atributos*

*Gols Marcados*

*Gols Sofridos*

*Número de Vitórias*

*Número de Derrotas*

*Número de Empates*

*Número de Jogos Disputados*

*(...)*



# Pré-Processamento

## Redução de Dimensionalidade

Tarefa que pode ser importante para questões de tempo de processamento, memória e, até mesmo, eficácia.



# Pré-Processamento

## Redução de Dimensionalidade



REDUÇÃO SIMPLES



TÉCNICAS AVANÇADAS

# Pré-Processamento

## Redução de Dimensionalidade

*Zhao and Cen (2013), por exemplo, demonstraram o uso da análise de componentes principais (PCA) para unir 13 variáveis que influenciam o resultado de uma partida de futebol, em uma única variável.*



# Pré-Processamento

## Seleção de Variáveis

*Outra forma de reduzir a dimensionalidade é através da seleção de um subconjunto de variáveis (feature selection).*



# Pré-Processamento

## Seleção de Variáveis



INTERNAS



FILTRO



ENVOLTÓRIO

# Pré-Processamento

## Discretização e Binarização

Outra forma de reduzir dimensionalidade é através da seleção de um subconjunto de variáveis, também conhecida como *feature selection*.



# Pré-Processamento

## Discretização e Binarização

Constatinou (2013), por exemplo, usou um sistema dinâmico [45] para discretizar variáveis como "força do time", "cansaço" e "motivação". A variável "cansaço", por exemplo, era determinada através da quantidade de dias entre um jogo e outro e, depois de aplicada a discretização, denotava valores como "baixa", "média" e "alta".

# Pesquisa Aplicada

## Processo KDD: Transformação de Dados

### Knowledge Discovery in Databases



# Transformação de Dados

## Conceito

Tarefa para adequar os dados à técnica de mineração a ser utilizada



# Transformação de Dados

## Métodos



A dark blue parallelogram containing the white, stylized mathematical symbol  $fx$ .

FUNÇÃO



NORMALIZAÇÃO



# Mineração de Dados

## Métodos



REGRESSÃO



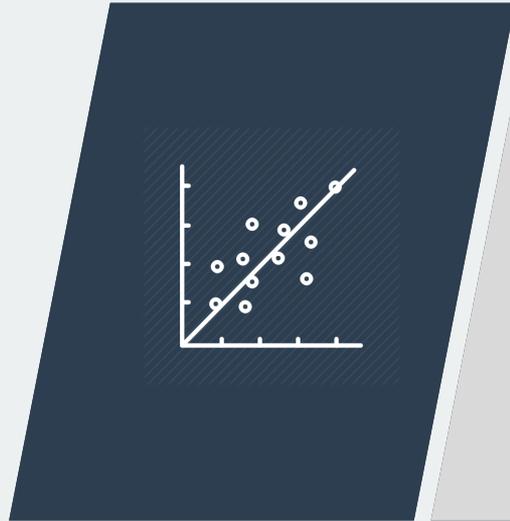
CLASSIFICAÇÃO



AGRUPAMENTO

# Mineração de Dados

## Métodos: Regressão



REGRESSÃO



CLASSIFICAÇÃO

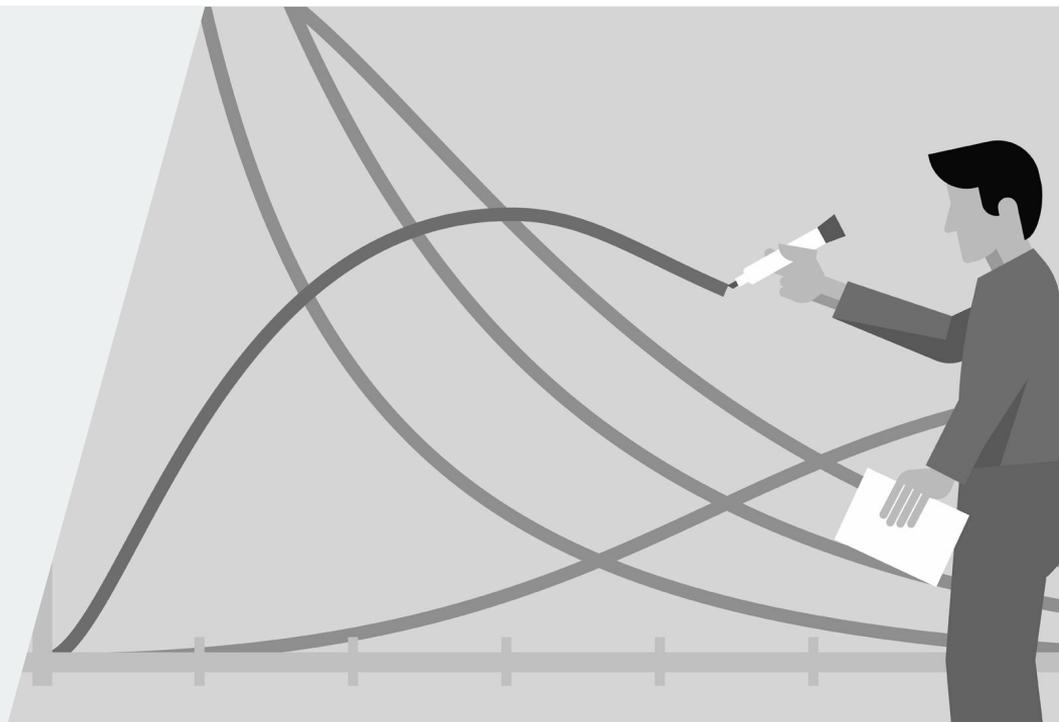


AGRUPAMENTO

# Regressão

## Conceito

É uma técnica de modelagem preditiva na qual a variável dependente (variável alvo) é contínua.



# Exemplos

## Regressão

Quantos **pontos** um time marcará num jogo de Basquete?

Quantas **horas** um tenista vai jogar para vencer um campeonato?

Por qual **preço** Cristiano Ronaldo deve ser negociado?

# Regressão Simples

## Regressão

$$y_i = w_0 + w_1 x_1 + \varepsilon_i$$

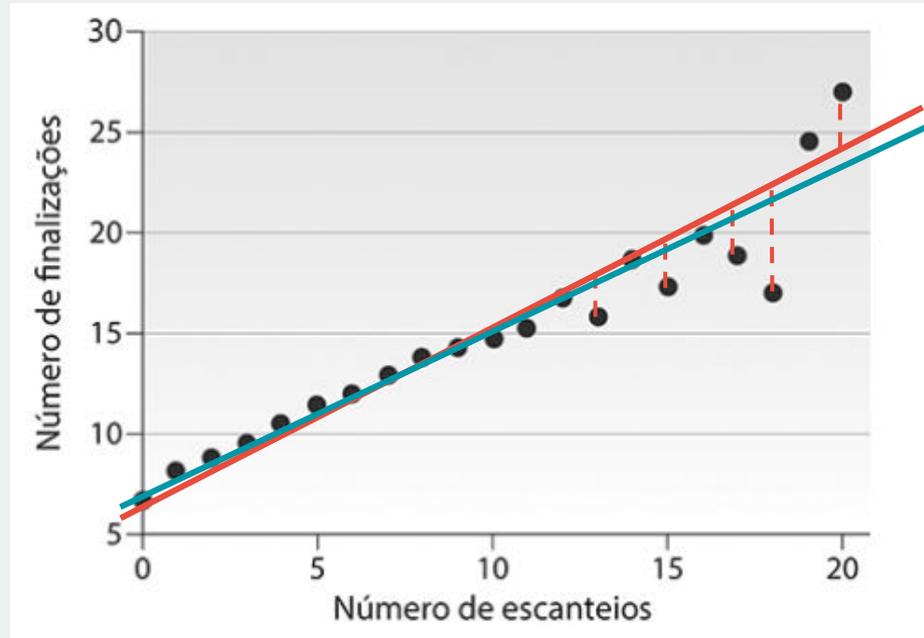
# Regressão Múltipla

## Regressão

$$\hat{y}(w, x) = w_0 + w_1x_1 + \dots w_px_p$$

# Mínimos Quadrados Ordinários

## Regressão



(adaptado de Anderson & Sally [30])

# Avaliação da Regressão

## Regressão

### *Soma dos Erros Quadrados*

$$RSS(w_0, w_1) = \sum_{i=1}^N (y_i - [w_0 + w_1 x_i])^2$$

# Métodos de Regressão

## Regressão

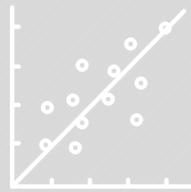
RIDGE  
REGRESSION

LASSO

NÃO  
PARAMÉTRICAS

# Mineração de Dados

## Métodos: Classificação



REGRESSÃO



CLASSIFICAÇÃO

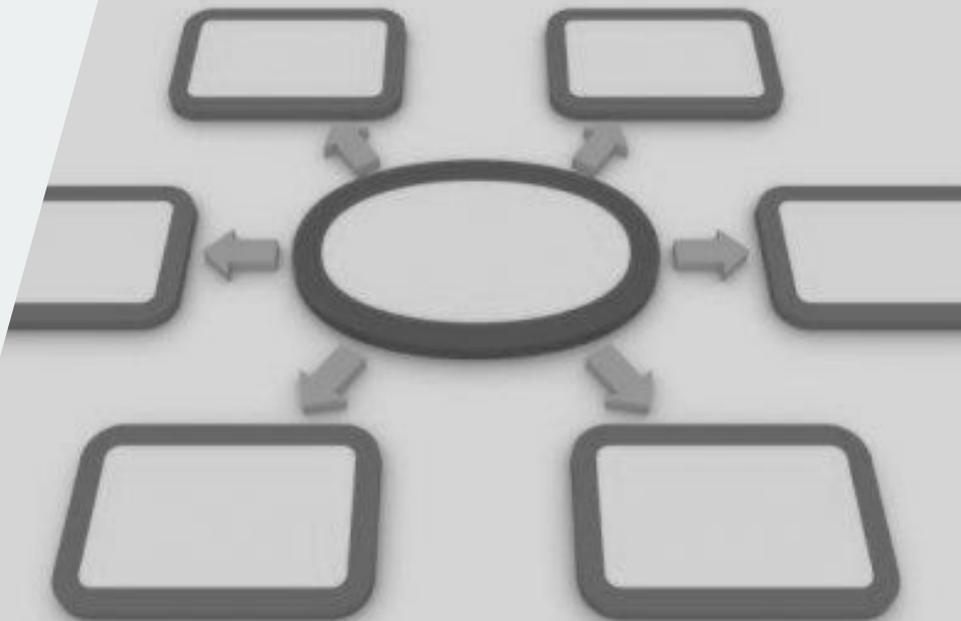


AGRUPAMENTO

# Classificação

## Conceito

É uma técnica de modelagem que mapeia objetos (ou observações) em uma das várias categorias pré-definidas.



# Exemplos

## Classificação

*Quem vencerá um jogo de vôlei?*

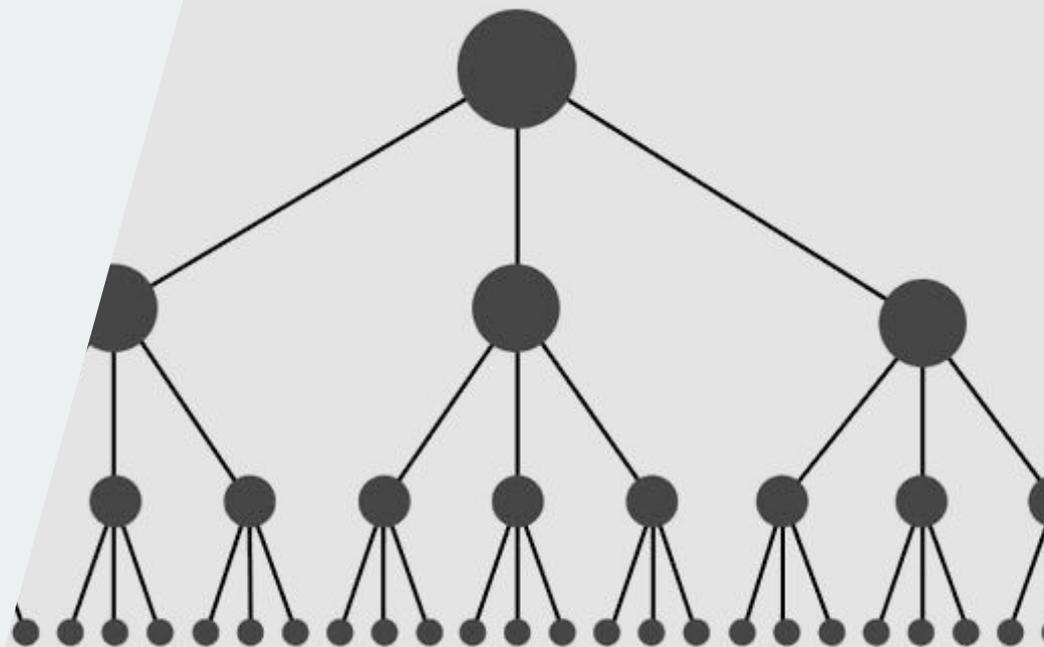
*Em uma partida de futebol, ambos os times irão marcar gols?*

*Qual golpe será o mais efetivo em uma luta de MMA?*

# Árvores de Decisão

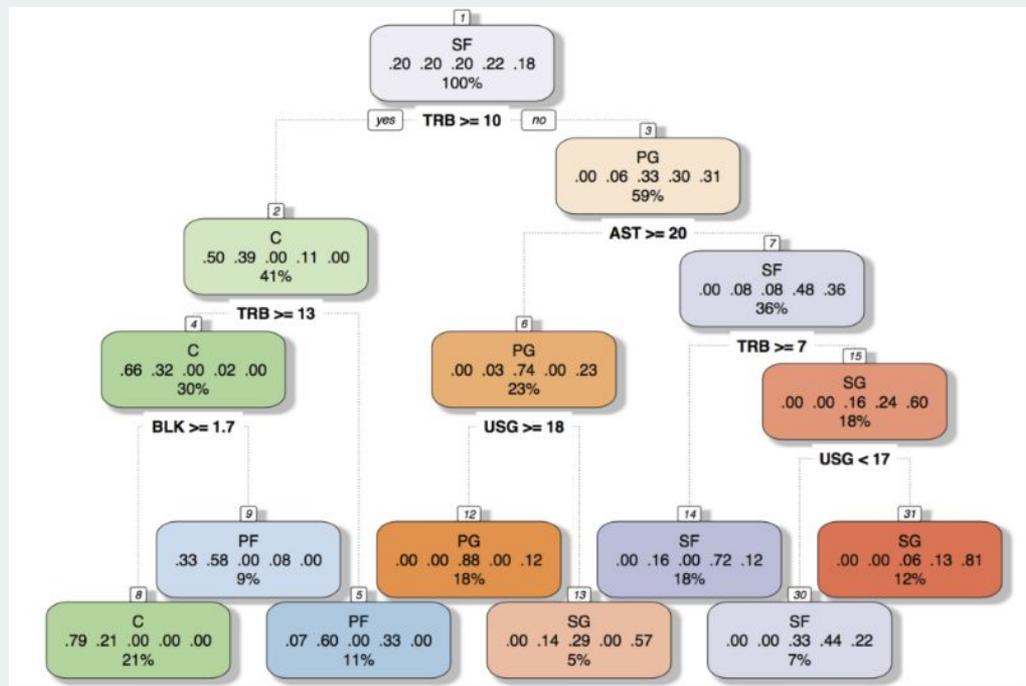
## Classificação

É uma estrutura semelhante a um fluxograma em que cada **nó interno** representa a avaliação de uma variável, cada **ramo** representa o resultado do teste e cada **nó de folha** representa um **rótulo de classe** (decisão tomada após o cálculo de todos os atributos). Os caminhos da raiz para a folha representam regras de classificação.



# Árvores de Decisão

## Classificação

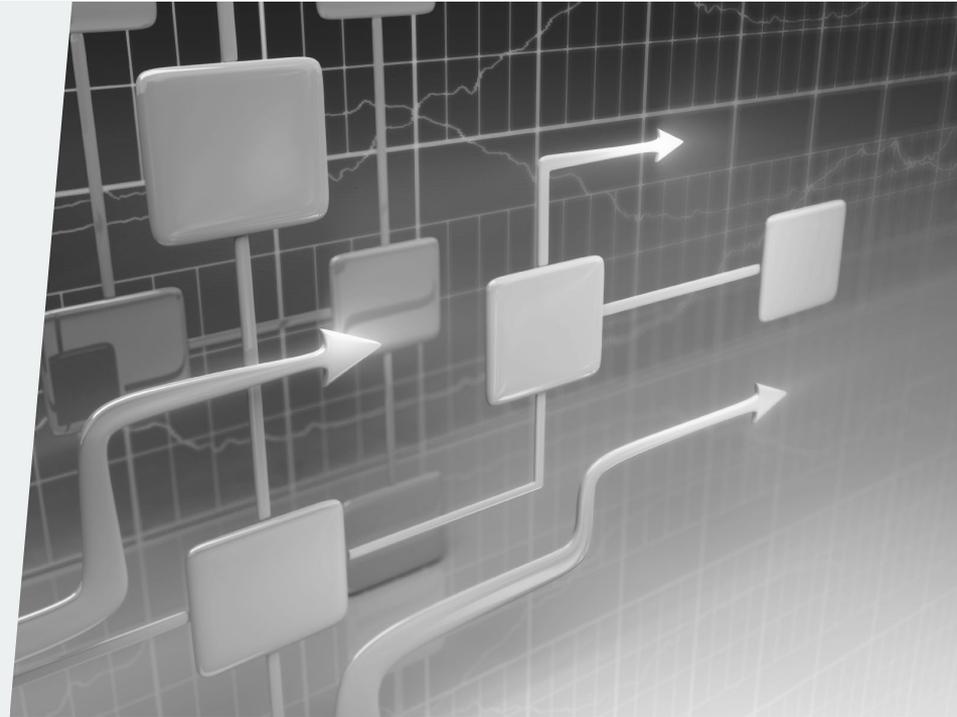


(adaptado de Jager [35])

# Redes Bayesianas

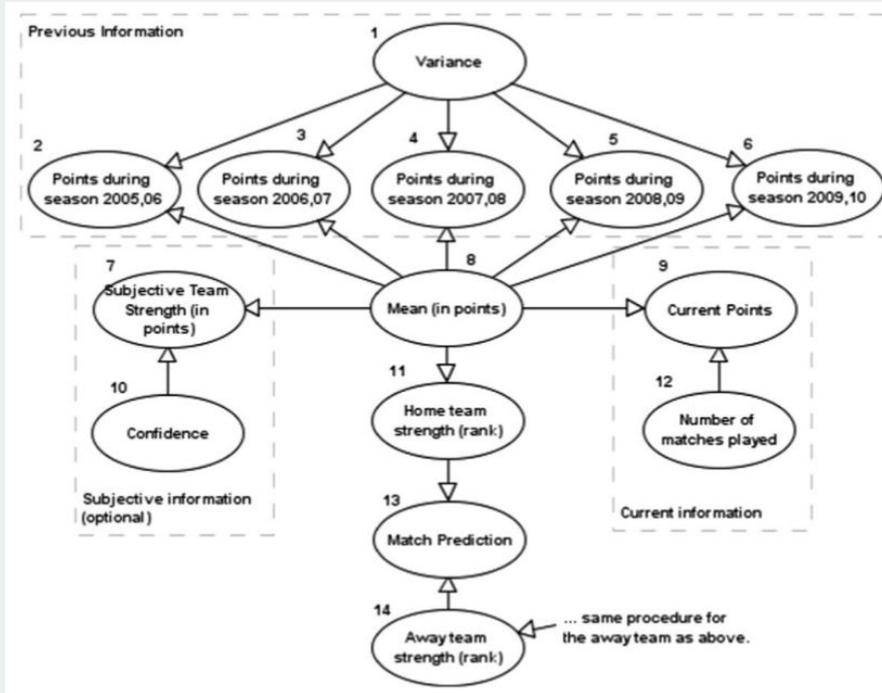
## Classificação

É um modelo acíclico e probabilístico que representa as variáveis e suas dependências através de um gráfico. Os nós representam as variáveis de um domínio, enquanto os arcos representam as dependências condicionais entre as variáveis. As informações sobre cada nó são dadas através da função de probabilidade que requer um determinado conjunto de valores como entrada e fornece uma distribuição de probabilidade de variáveis como saída



# Redes Bayesianas

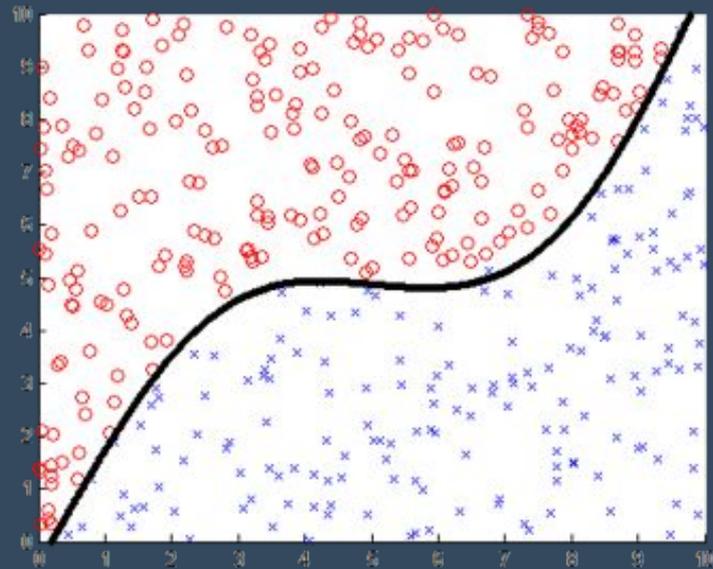
## Classificação



# Máquinas de Vetor de Suporte

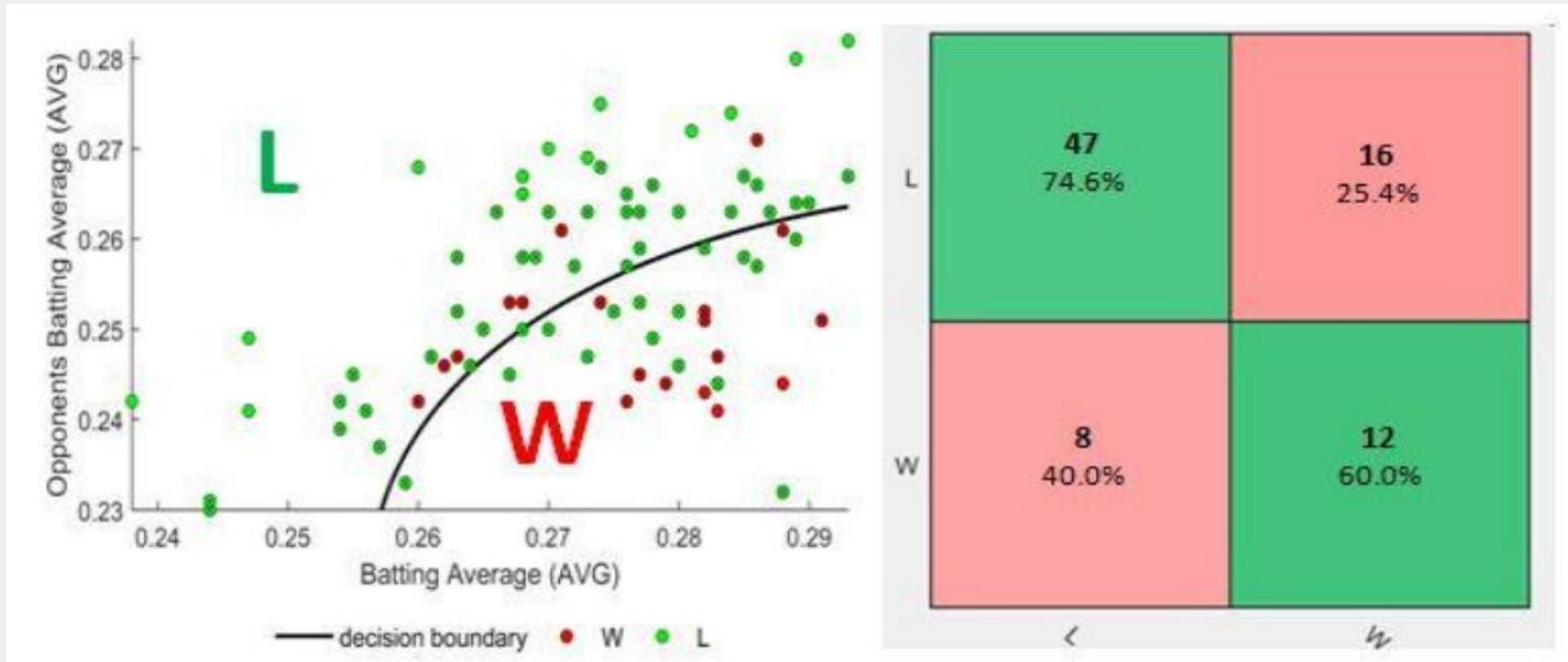
## Classificação

A Máquina de Vetor de Suporte (MVS) analisa, para cada observação de conjunto de dados, qual de duas possíveis classes a observação faz parte.



# Máquinas de Vetor de Suporte

## Classificação



(adaptado de Tolbert & Trafalis [50])

# Redes Neurais

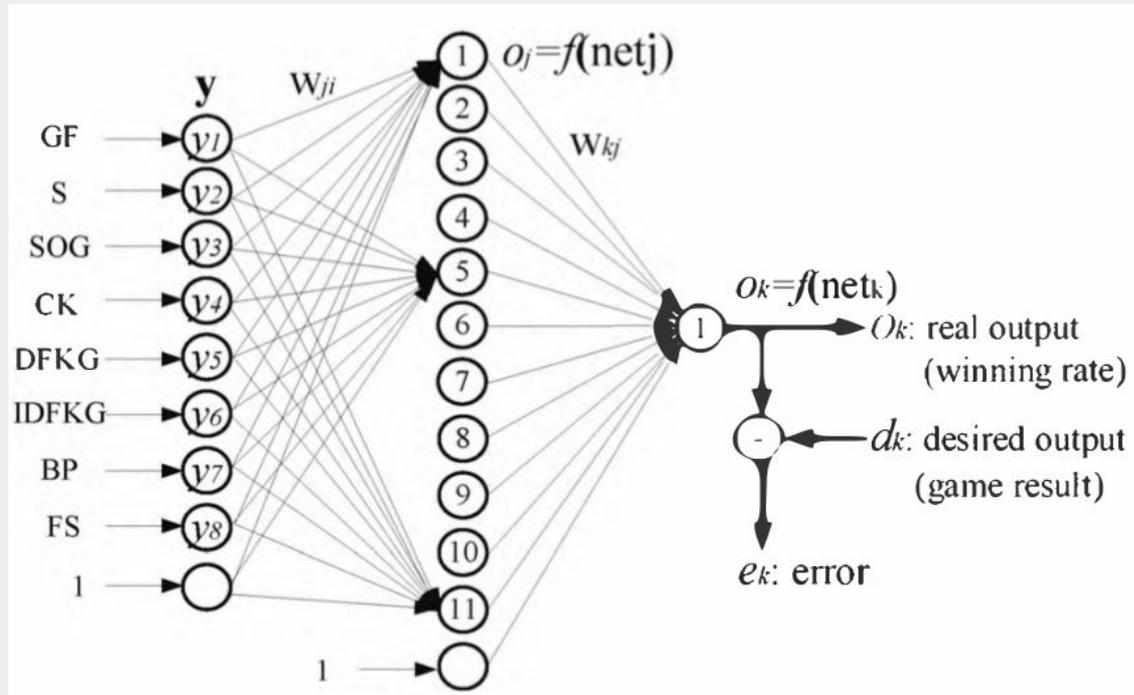
## Classificação

Redes Neurais Artificiais são técnicas computacionais que apresentam um modelo matemático inspirado na estrutura neural de organismos inteligentes e que adquirem conhecimento através da experiência



# Redes Neurais

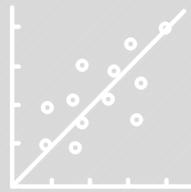
## Classificação



(adaptado de Huang [35])

# Mineração de Dados

## Métodos: Agrupamento



REGRESSÃO



CLASSIFICAÇÃO



AGRUPAMENTO

# Conceito

## Agrupamento

É uma técnica para agrupar observações segundo algum grau de semelhança de forma automática (sem intervenção humana)

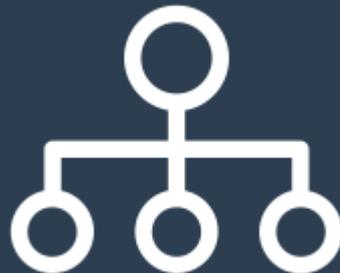


# Tipos

## Agrupamento



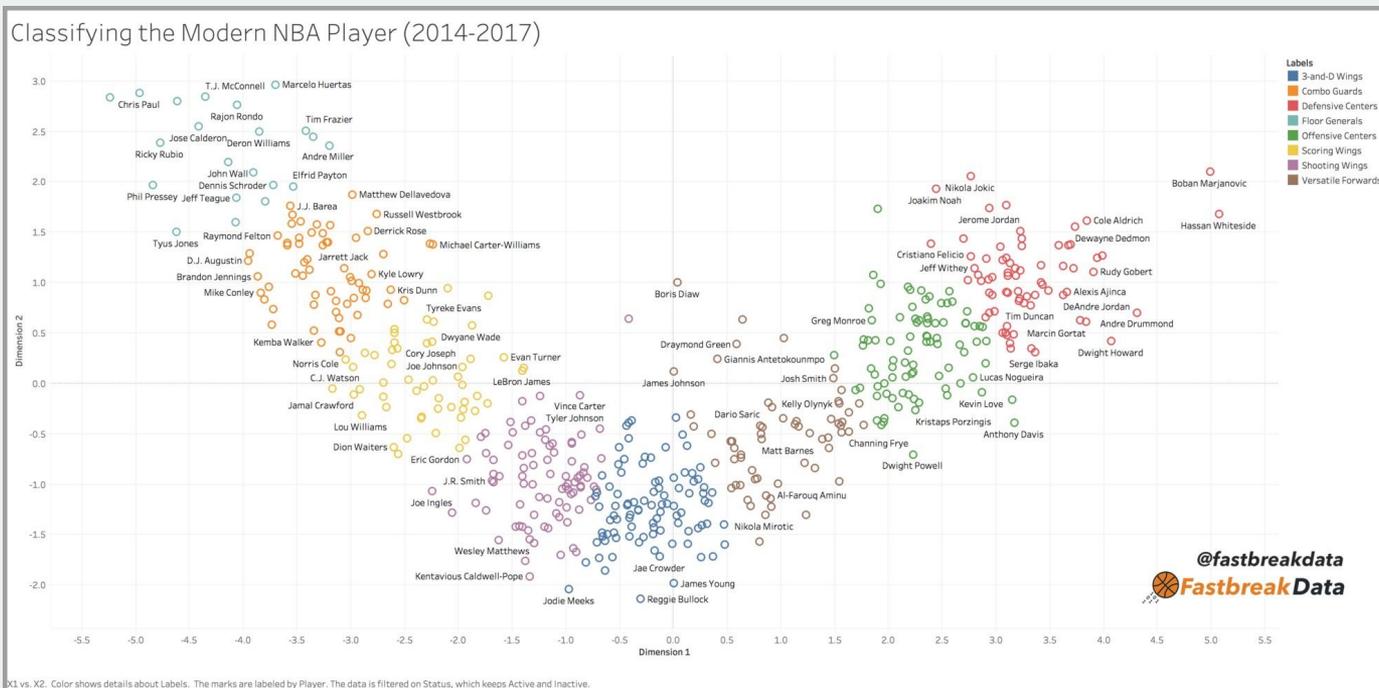
BASEADO EM PARTIÇÃO



BASEADO EM HIERARQUIA

# Exemplo: Partição

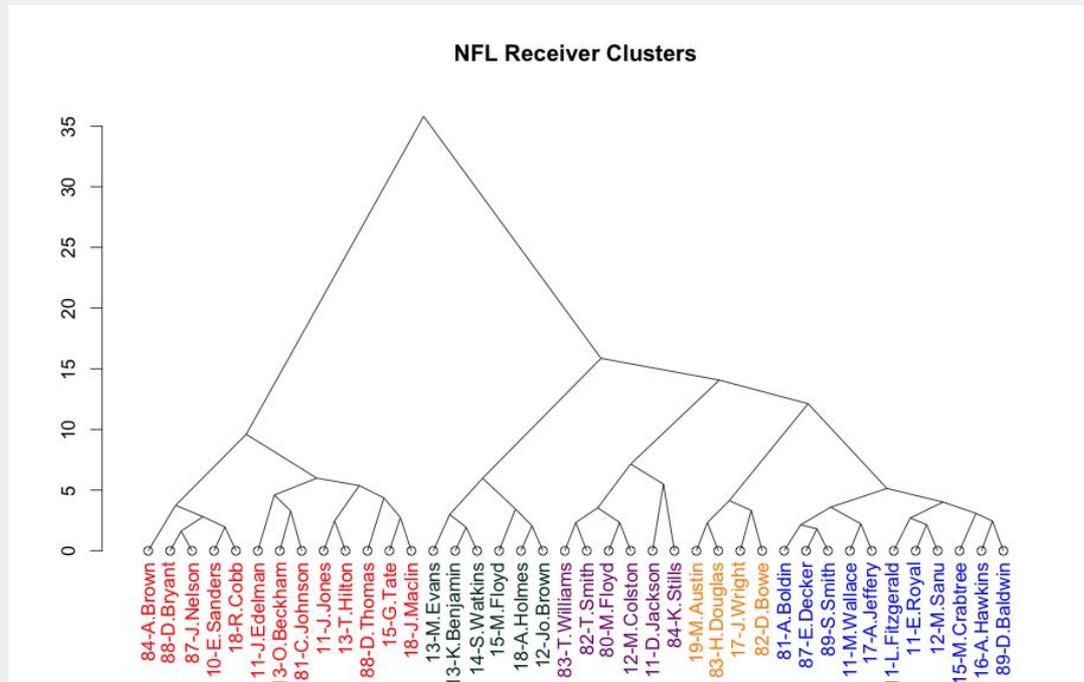
## Agrupamento



(adaptado de Cheng [27])

# Exemplo: Hierárquico

## Agrupamento

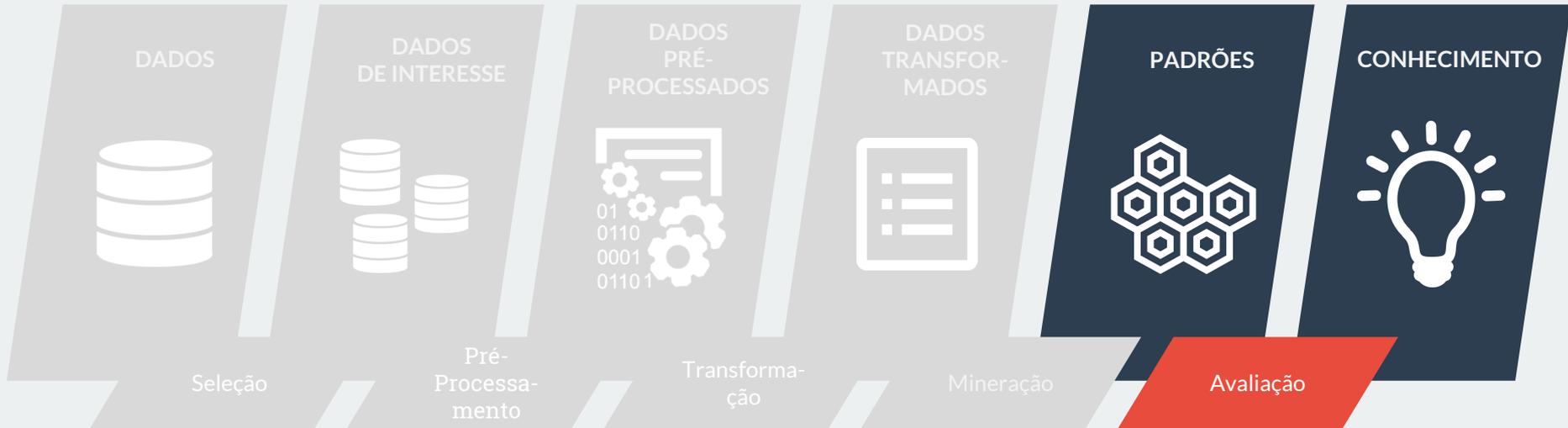


(adaptado de Lesmeister [2])

# Pesquisa Aplicada

## Processo KDD: Avaliação e Interpretação

### Knowledge Discovery in Databases

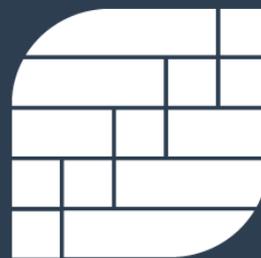


# Avaliação e Interpretação

## Métodos



HOLDOUT



VALIDAÇÃO CRUZADA

# Avaliação e Interpretação

## Holdout



TREINAMENTO E  
VALIDAÇÃO



TESTE

# Avaliação e Interpretação

## Validação Cruzada K-Fold



# Revisão

## Processo de KDD



# Pesquisa Aplicada

**Prática: 30 minutos**



<https://github.com/igormago/sbbd17>

<https://try.jupyter.org/>

# Prática

## Linguagens Mais Usadas



R



PYTHON

# Prática

## Ferramentas e Bibliotecas



ANACONDA®

ANACONDA

pandas

$$y_i t = \beta' x_{it} + \mu_i + \epsilon_{it}$$



PANDAS



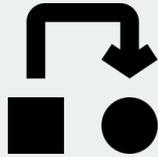
SCIKIT LEARN



JUPYTER

# Desafios Emergentes

## Pesquisa Aplicada



*Transformar dados de  
movimentos em dados de  
eventos*

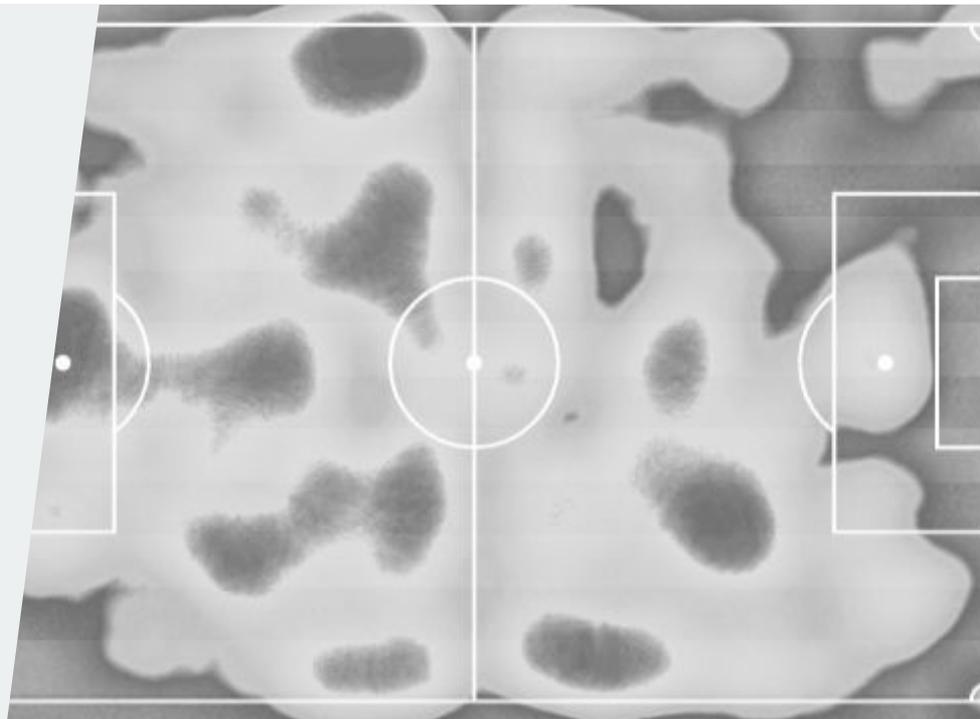


# Desafios Emergentes

## Pesquisa Aplicada



Prever *onde* os eventos vão  
acontecer

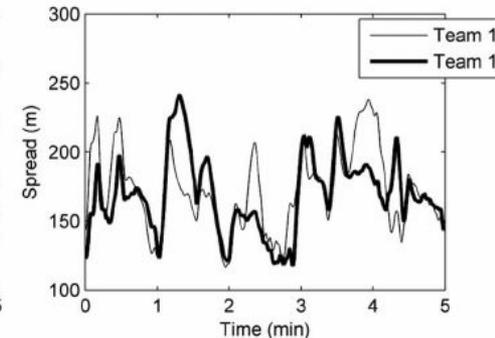
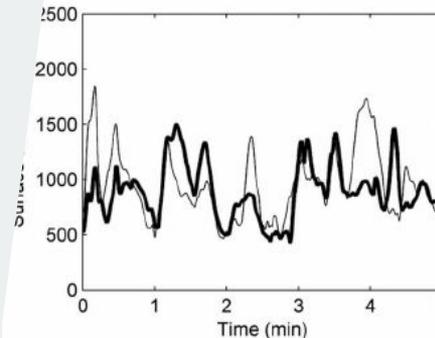
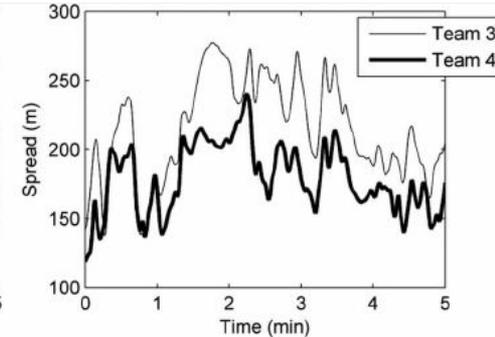
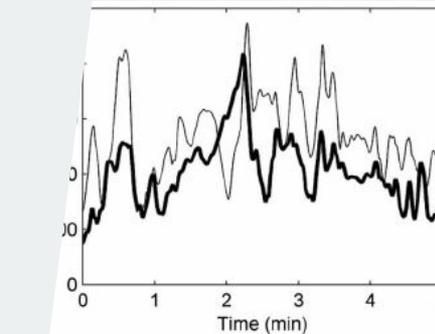


# Desafios Emergentes

## Pesquisa Aplicada



Prever *quando* os eventos vão acontecer

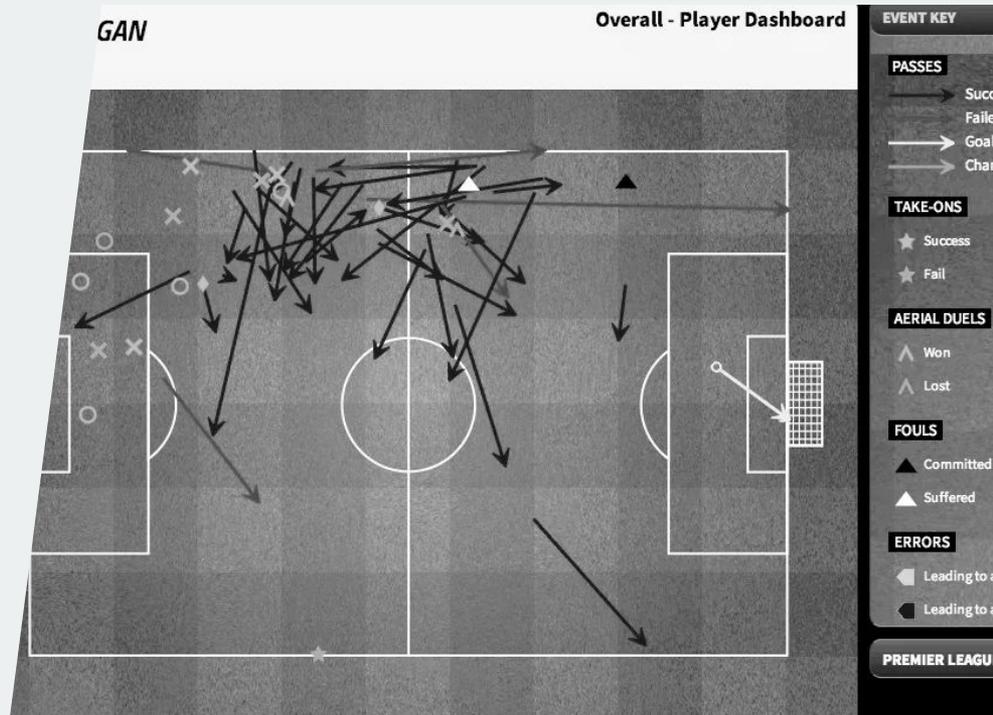


# Desafios Emergentes

## Pesquisa Aplicada



Analisar o **desempenho** dos atletas



# Desafios Emergentes

## Pesquisa Aplicada



Prever **resultados** de jogos,  
campeonatos ou quantidade  
de eventos



# Desafios Emergentes

## Pesquisa Aplicada



Resolver desafios específicos  
de Sports **Betting** Analytics





**04**

**Apostas  
Esportivas**

# Apostas Esportivas

## Contexto

*Mercado avaliado em 3 trilhões de dólares que representa 37% do mercado de “jogos de azar”*



# Apostas Esportivas

## Contexto

*Crescimento das casas de apostas online hospedadas em países onde o jogo é regulamentado*



# Conceitos Fundamentais

## Tipos de Mercado



CASA DE APOSTAS

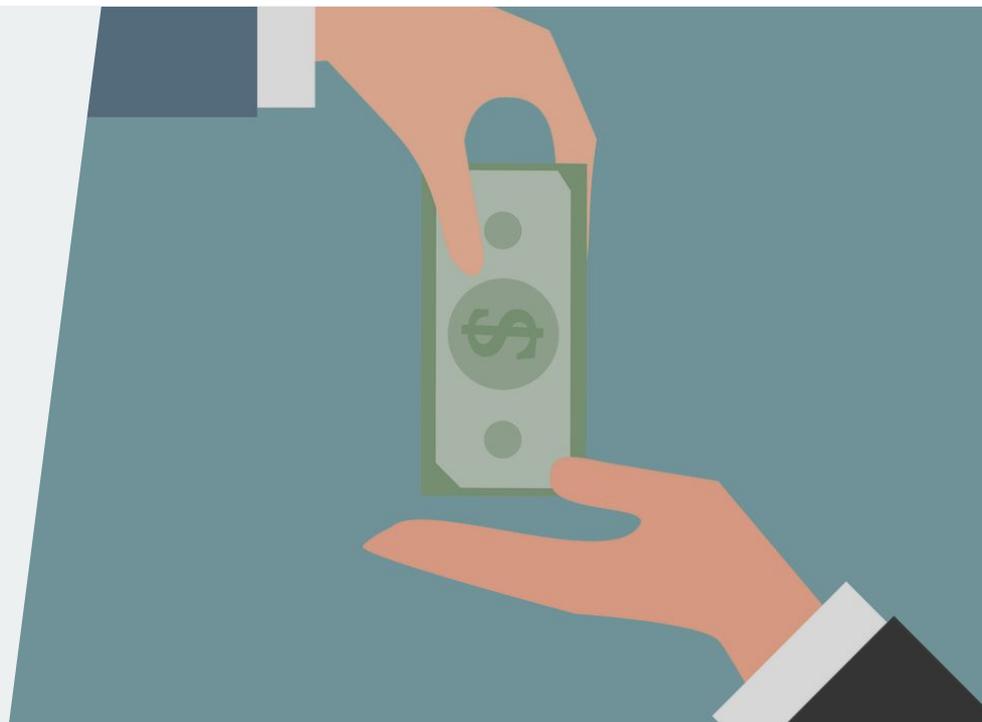


BOLSA DE APOSTAS

# Conceitos Fundamentais

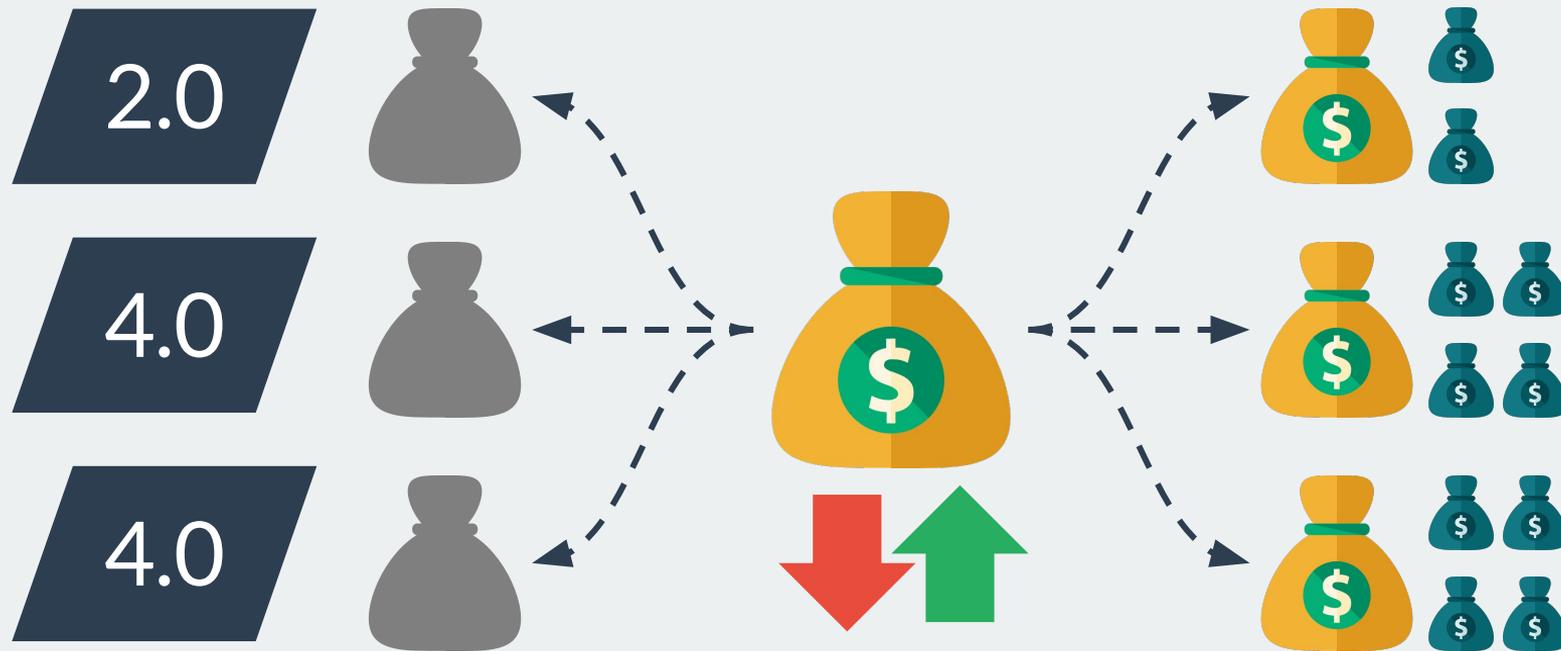
## Odds

*Define o quanto um apostador receberá, se fizer uma aposta bem sucedida.*



# Conceitos Fundamentais

## Odds



# Conceitos Fundamentais

## Odds



# Conceitos Fundamentais

## Probabilidade Implícita

*Toda odd representa  
**implicitamente** a  
probabilidade de um evento  
ocorrer*



# Conceitos Fundamentais

## Probabilidade Implícita



**vs.**



$$100/2.0 = 50\%$$

$$100/4.0 = 25\%$$

$$100/4.0 = 25\%$$

# Conceitos Fundamentais

## Overround



# Conceitos Fundamentais

## Overround



# Conceitos Fundamentais

## Overround x Lucro das Casas de Apostas



vs.



R\$50\*1.9  
R\$95.00

+5%

R\$25\*3.8  
R\$95.00

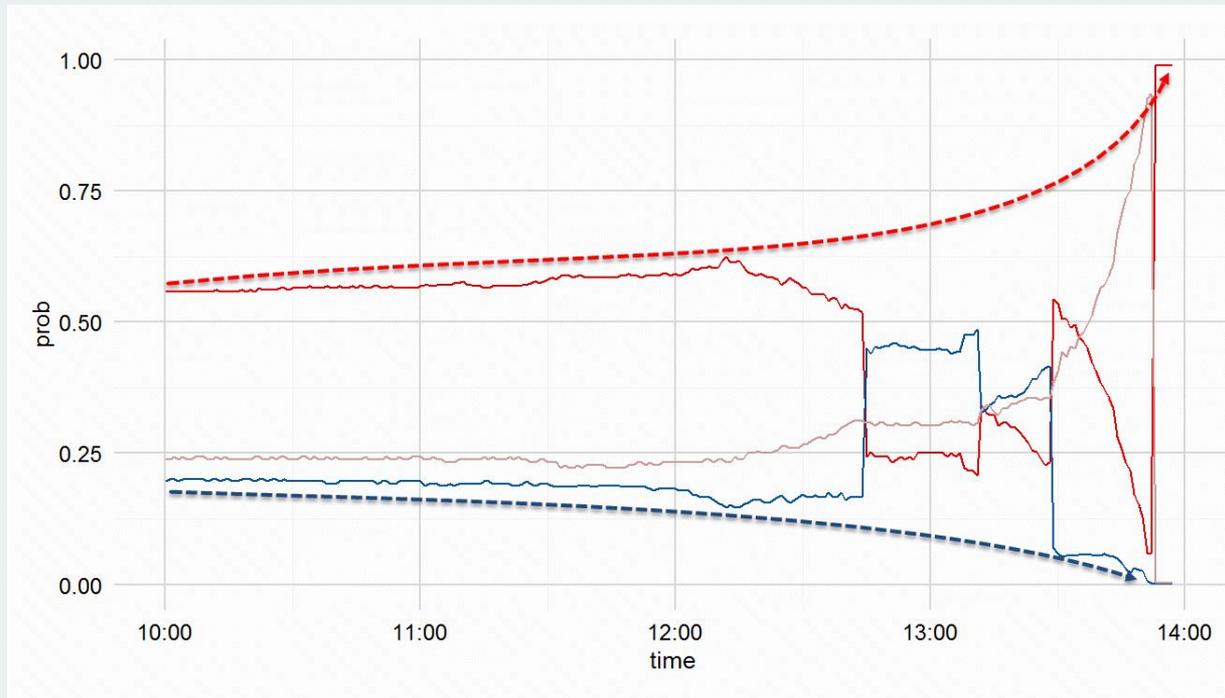
+5%

R\$25\*3.8  
R\$95.00

+5%

# Apostas Esportivas

## Varição das Odds



# Apostas Esportivas

## Desafios Emergentes



*Avaliação de estratégias de  
gestão de banca*



# Apostas Esportivas

## Desafios Emergentes



*Avaliação de eficiência de mercado*



# Apostas Esportivas

## Desafios Emergentes



*Avaliação e modelagem de estratégias para trading.*

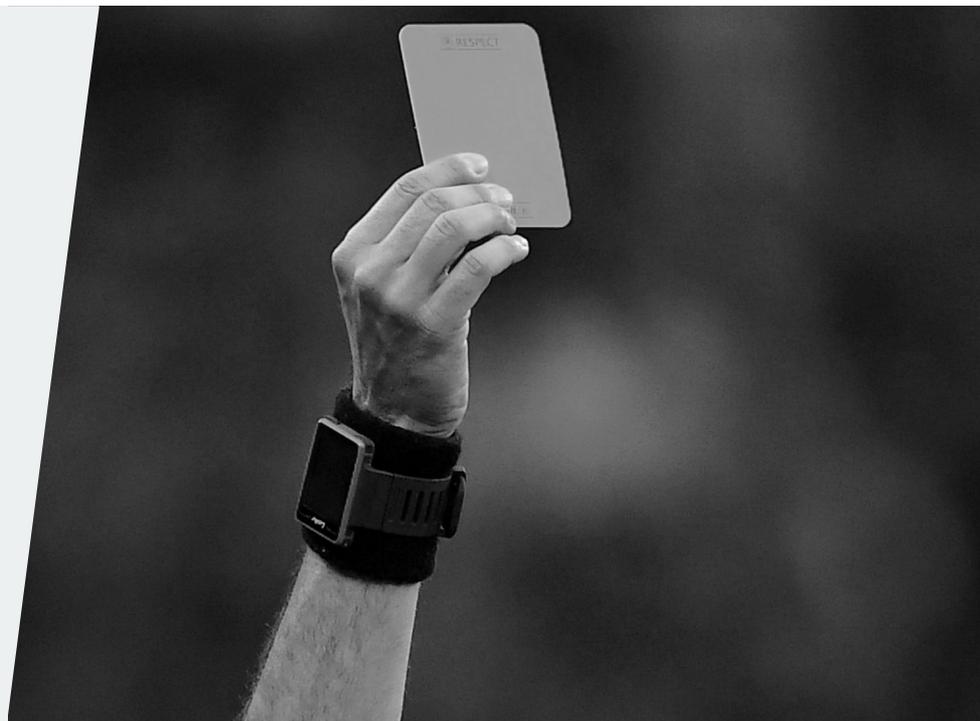


# Apostas Esportivas

## Desafios Emergentes



*Detecção de fraudes (jogos manipulados).*



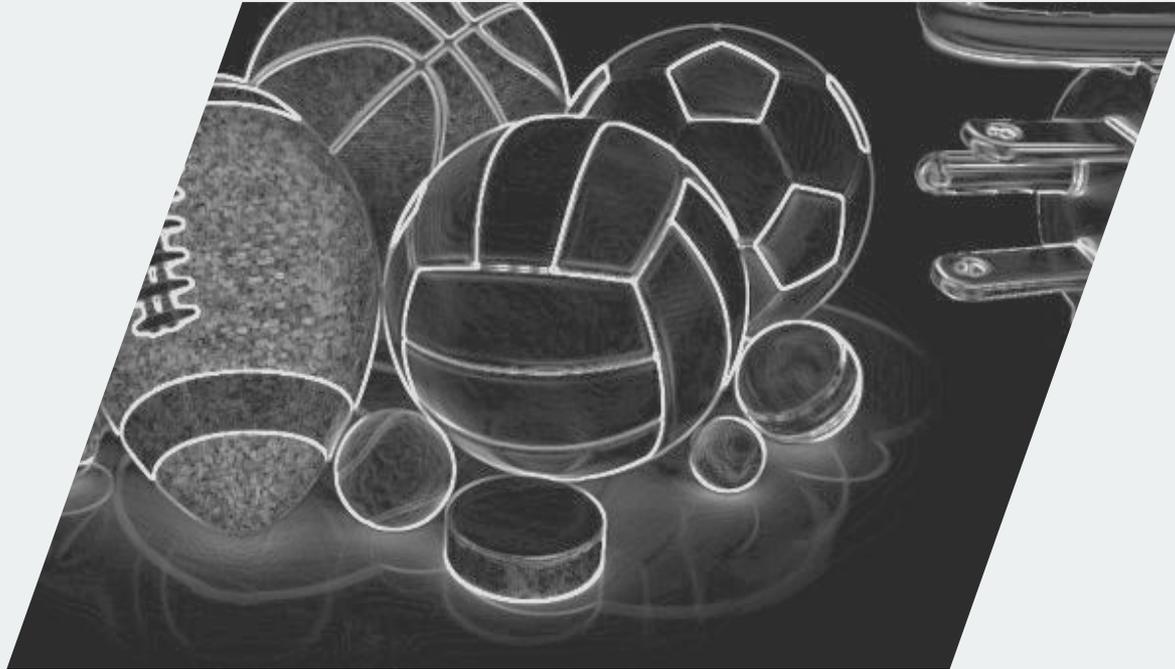
05

# Considerações Finais



# O que será do futuro?

## Considerações Finais



05

# Considerações Finais



**OBRIGADO!**



# Contato

## Sports Analytics - Mudando o Jogo



igor.costa@ifpb.edu.br

cesp@dsc.ufcg.edu.br

lbmarinho@dsc.ufcg.edu.br