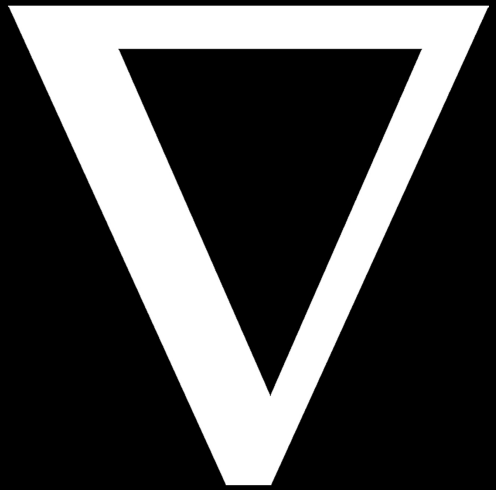


Introdução à Ciência de Dados com Python

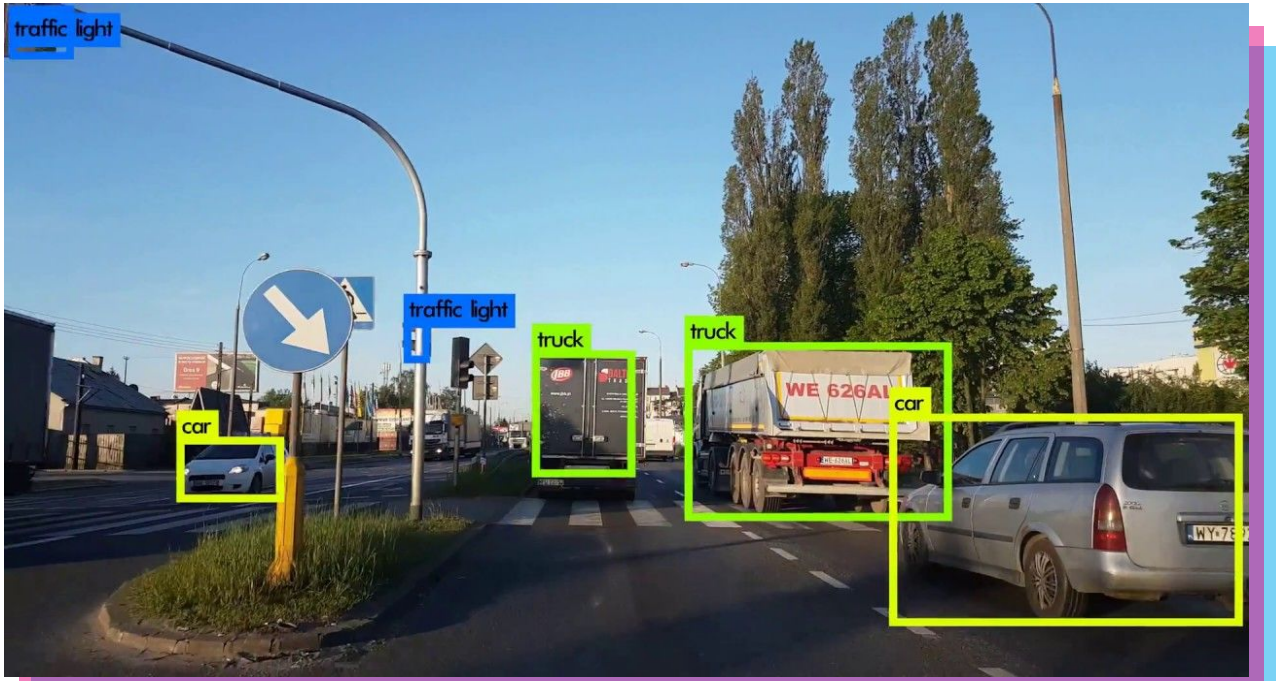
Gustavo Sutter
@suttergustavo

Apresentação: quem somos

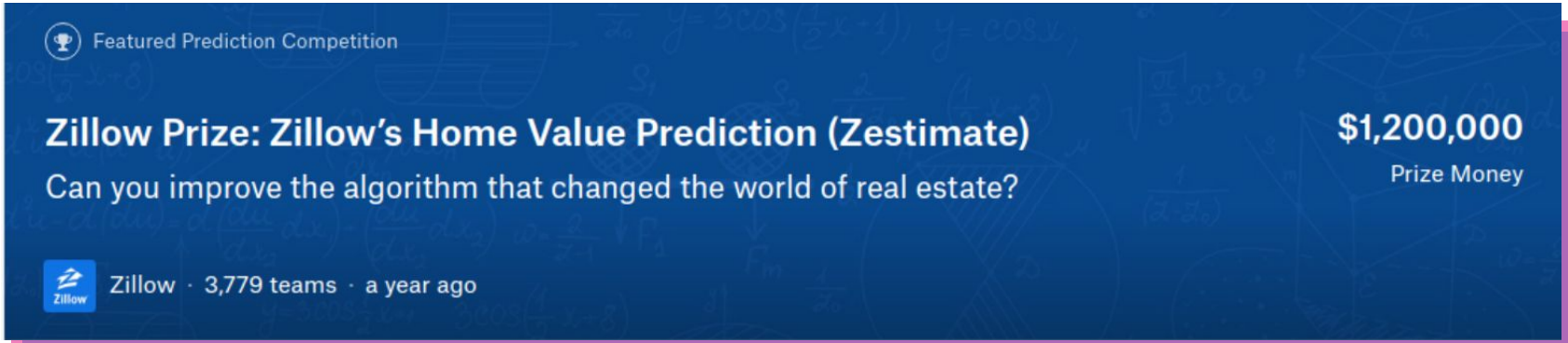
- Somos um grupo de extensão do Instituto de Ciências Matemáticas e de Computação da USP, em São Carlos
- Nosso objetivo é popularizar ciência de dados e aprendizado de máquina através de cursos, palestras e eventos
- Para mais informações acesse nosso site
    - http://data.icmc.usp.br/

Apresentação: o que você vai aprender

- Fundamentos teóricos e práticos da ciência de dados usando Python
- Utilizar Jupyter Notebook como ferramenta de trabalho
- Funcionamento das principais bibliotecas utilizadas tanto na academia quanto na indústria
    - NumPy, Pandas, Matplotlib, Scikit-Learn
- Todo material (códigos e slides) pode ser encontrado em nosso GitHub:
    - https://github.com/icmc-data/Intro-Ciencia-de-Dados-Youtube



Classificação e localização de objetos em imagens (mais...)



Predição do preço de imóveis (mais...)

Figure 1: Class-conditional samples generated by our model.

Geração de imagens artificiais (mais...)

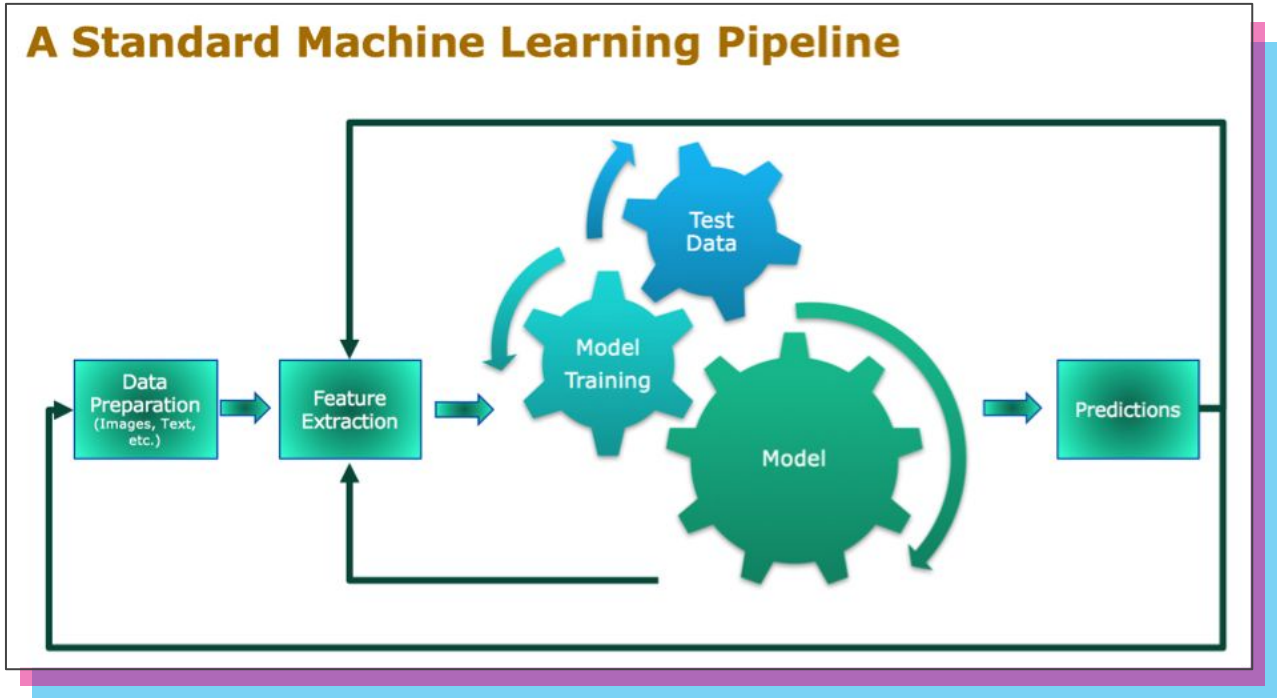
Jogar (e ganhar) contra humanos em jogos complexos (mais...)

O que é o que é?

- Artificial Intelligence
  - Sistemas que simulam comportamento humano (dele alguma forma).
- Machine Learning
  - Encontrar padrões em dados para melhorar performance em alguma tarefa.
- Deep Learning
  - Categoria específica de ML, que usa redes neurais profundas.
- Data Science
  - Todo o ***pipeline***, desde o pré-processamento (ou até coleta) ao uso prático do modelo.
- Data Engineering
  - Área que dá o suporte de infraestrutura para as tarefas de data science.

Etapas do processo

- Obtenção dos dados
- Análise dos dados
- Pré processamento
- Escolha e treinamento do modelo
- Análise dos resultados
- **repeat...**



Pipeline de DS/ML (fonte)

Dados

- Dados estruturados
  - Tabelas (Geralmente chamamos de dados tabelados)
- Dados não estruturados
  - Imagens, textos, áudio, …
- Mas eles podem aparecer combinados
  - Por exemplo: Uma tabela sobre vendas pode ter que alguns campos são textos de reviews

Dados tabulados

Colunas -> Atributos/Features

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Linhas -> Exemplos/Instâncias

Dados da base Titanic (link)

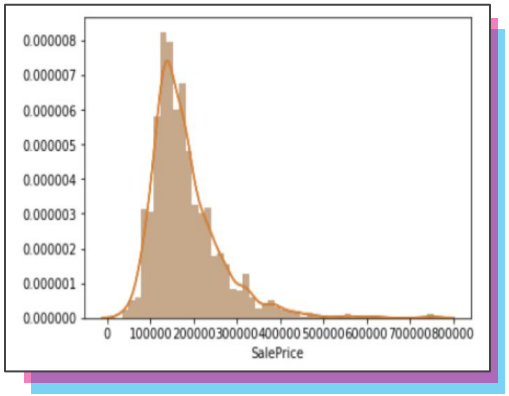
Tipos dos atributos

- De forma simplificada um atributo pode ser de um entre em dois tipos:
  - Numérico (Por exemplo: peso, idade, altura, área, preço)
  - Categórico (Por exemplo: cidade, sexo, modelo)
- Depois vamos ver que cada um desses grupos se divide em dois subgrupos mais específicos
- O tipo do atributo é de extrema importância, pois pode ser aceito ou não por um modelo e a conversão entre tipos afeta diretamente seu funcionamento

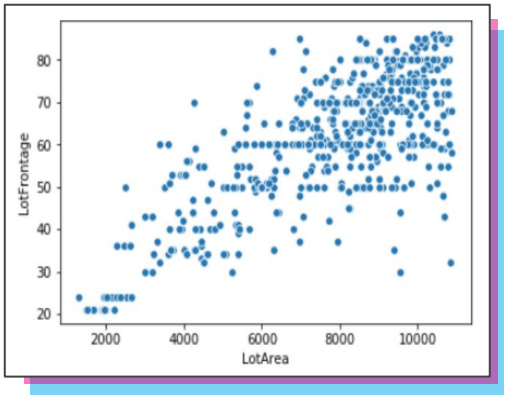
Análise exploratória (Exploratory Data Analysis -EDA)

- Etapa que serve para a **compreensão dos dados**
- Fazemos observações estatísticas sobre as features
    - Min/Max
    - Média
    - Mediana
    - Desvio Padrão e Variância
- Verificação da consistência e qualidade dos dados (e.g., valores nulos)
- **Utilização de gráficos** para melhor entendimento dos atributos

Alguns dos gráficos utilizados



Histograma



Scatter plot

Pré processamento

- Nem sempre (quase nunca) a base vem limpa e pronta para ser colocada no modelo
- Muitas vezes encontramos coisas estranhas:
  - Valores nulos (e.g., a pessoa não preencheu sua idade em um cadastro)
  - Valores inconsistentes (e.g., produto não vendido com nome do comprador)
  - Valores inválidos (e.g., altura com valor negativo)
- Feature engineering (geração de novos atributos para enriquecer a base)
  - Vamos supor que temos peso e altura das pessoas em nossa base, podemos usar essas informações para calcular seu IMC, o que pode melhorar o funcionamento do modelo

Está bom de introdução por enquanto

- Até aqui vimos que antes de começar a pensar em utilizar um modelo para realizar predições temos que fazer muita coisa
- No próximo vídeo iremos começar esse processo de manipulação de dados na prática, entendendo como podemos fazer cada uma dessas coisas usando Python.