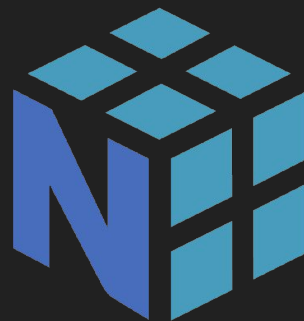




# Iniciando em Ciência de Dados com Python



Lucas Molitor





Contexto



# BIG DATA

- Paradigma caracterizado por grandes conjuntos de dados, que aplicativos tradicionais não são capazes de lidar;
- 3V's: Volume, Variedade e Velocidade;
- Valor e Veracidade;



# TENDÊNCIAS DA INDÚSTRIA 4.0

- Manufatura Digital;
- Automação Industrial;
- IoT;
- IA;
- Cloud Computing;
- Realidade Aumentada;
- Produção Aditiva (Impressão 3D);
- Segurança da Informação;
- Big Data;





O que é um dado?



# DADO

- Unidade indivisível, objetiva e geralmente abundante, com o papel de registrar um fato;
- Menor e mais simples elemento de um sistema;
- Fácil manipulação e transporte;



# INFORMAÇÃO e CONHECIMENTO



- Informação é um conjunto de dados adicionados a um contexto;
- Conhecimento é aquilo que quando adquirido, através do tratamento das informações, é capaz de mudar o comportamento de um sistema;




E a Ciência de Dados?





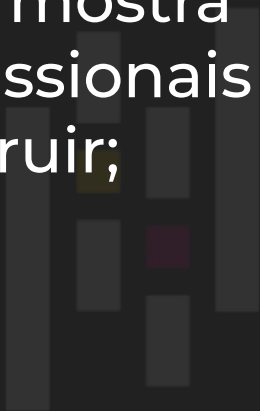

# CIÊNCIA DE DADOS



- Estudo e análise de dados visando:
    - Extrair conhecimento;
    - Detectar padrões;
    - Obter Insights para tomadas de decisão;
  - Onde a Programação, Tecnologia, Matemática e Estatística se encontram;
- 

# CIÊNCIA DE DADOS



- Exige, além das habilidades de programação e/ou estatísticas, conhecimento de setor;
  - Se mostra uma área interdisciplinar na qual profissionais dos mais variados setores podem atuar e usufruir;
- 
- 

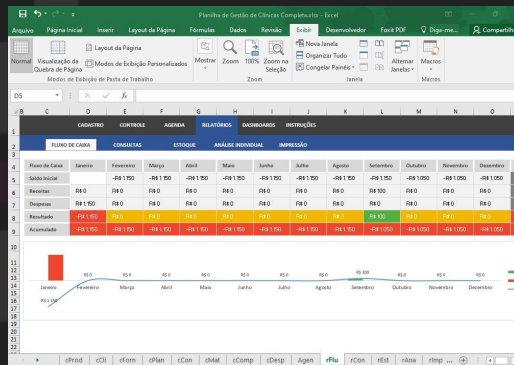


Que tipos de dados podemos encontrar?



# DADOS ESTRUTURADOS

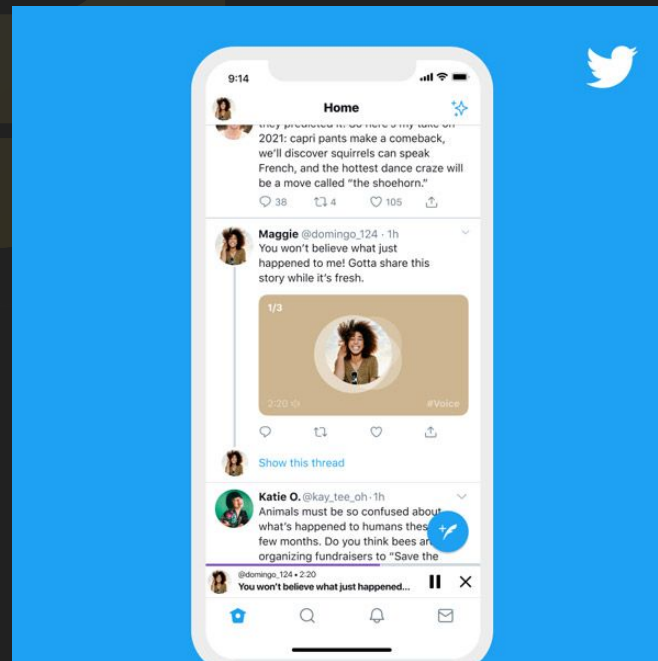
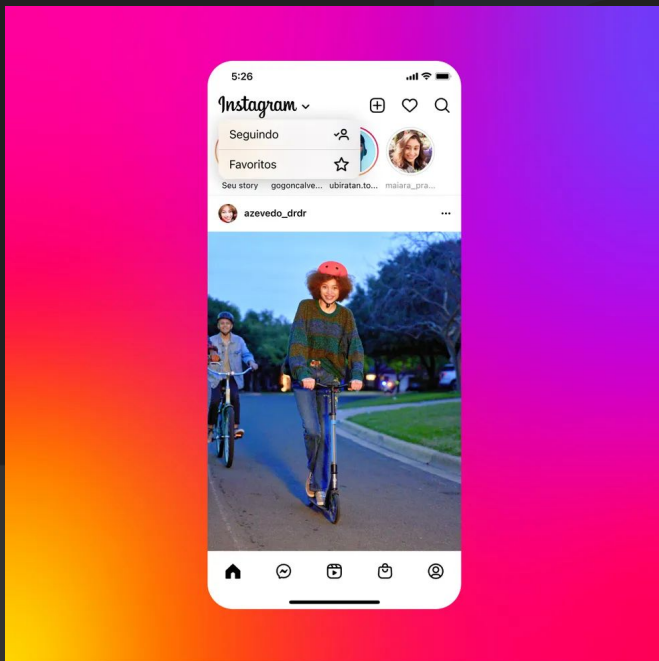
- Formato e comprimento definido;
- Dados armazenados em bancos de dados, planilhas, etc;



# DADOS NÃO ESTRUTURADOS

- Formato e comprimento indefinido;
- Não seguem um padrão, são flexíveis;
- Dados de texto, áudio, vídeos, imagens, etc;
- A maior massa de dados existente na atualidade, graças às redes sociais;

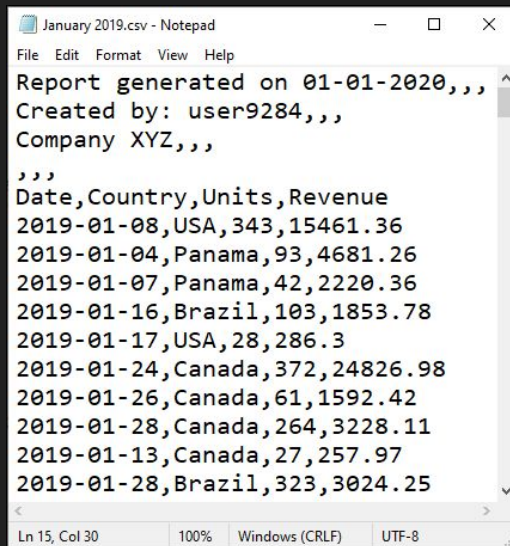
# DADOS NÃO ESTRUTURADOS



Conteúdos do feed do Instagram e Twitter como exemplos de dados não estruturados

# DADOS SEMI-ESTRUTURADOS

- Formato e comprimento parcialmente definido;
- Estrutura implícita e flexível;
- Arquivos CSV, XML, JSON;



```
January 2019.csv - Notepad
File Edit Format View Help
Report generated on 01-01-2020,,,
Created by: user9284,,,
Company XYZ,,,

'''
Date,Country,Units,Revenue
2019-01-08,USA,343,15461.36
2019-01-04,Panama,93,4681.26
2019-01-07,Panama,42,2220.36
2019-01-16,Brazil,103,1853.78
2019-01-17,USA,28,286.3
2019-01-24,Canada,372,24826.98
2019-01-26,Canada,61,1592.42
2019-01-28,Canada,264,3228.11
2019-01-13,Canada,27,257.97
2019-01-28,Brazil,323,3024.25
```

# DADOS SEMI-ESTRUTURADOS

```
{  
  "nome": "João da Silva",  
  "idade": 35,  
  "cidade": "São Paulo",  
  "telefone": "(11) 1234-5678",  
  "email": "joao.silva@email.com"  
}
```

Arquivo JSON

```
▼<Servicos>  
  ▼<cServico>  
    <Codigo>04014</Codigo>  
    <Valor>62,09</Valor>  
    <PrazoEntrega>6</PrazoEntrega>  
    <ValorSemAdicionais>22,50</ValorSemAdicionais>  
    <ValorMaoPropria>0,00</ValorMaoPropria>  
    <ValorAvisoRecebimento>0,00</ValorAvisoRecebimento>  
    <ValorValorDeclarado>39,59</ValorValorDeclarado>  
    <EntregaDomiciliar>S</EntregaDomiciliar>  
    <EntregaSabado>S</EntregaSabado>  
    <obsFim/>  
    <Erro>0</Erro>  
    <MsgErro/>  
  </cServico>  
</Servicos>
```

Arquivo XML



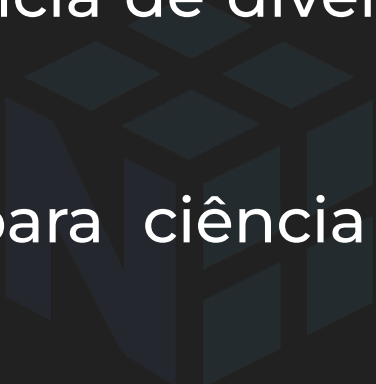


Onde o Python entra nisso tudo?



# PYTHON

The Python logo, consisting of two interlocking snakes, one blue and one yellow, is positioned in the upper center of the slide.

- Linguagem de programação Open Source;
  - Simples e poderosa;
  - Por ser Open Source, permite a existência de diversas bibliotecas;
  - Razão principal pelo qual é usado para ciência de dados;
- 
- A 3D graphic of a cube, composed of smaller cubes, is located in the bottom right corner of the slide.

# NumPy e Pandas

- NumPy é uma biblioteca para manipulação de conjuntos de dados numéricos e operações matemáticas;
- Pandas é uma biblioteca para ciência de dados, capaz de otimizar operações e fornecer funcionalidades prontas para tratamento dos mesmos, agilizando o processo;

# NumPy e Pandas

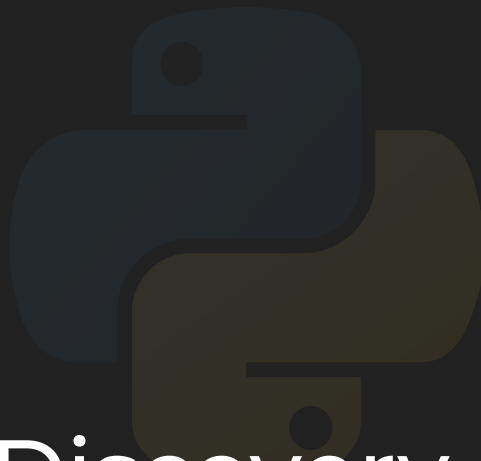
- Não são as únicas bibliotecas! Entretanto são as mais utilizadas por serem simples e muito eficientes. Entretanto apesar de poderosas, assim como tudo na programação, podem ter suas limitações dependendo da sua necessidade;





Chega de teoria por enquanto





# Knowledge Discovery in Databases (KDD)



# KDD

- Extração de Conhecimentos;
- Processo de várias etapas não trivial, iterativo e iterativo, para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis a partir de um conjunto de dados;



# FASES DO KDD



## Seleção

Selecionar o conjunto de dados que fará parte da análise.



## Pré-Processamento

Verificar a qualidade dos dados, tratar possíveis ruídos.



## Transformação

Aplicar técnicas de transformação: normalização, agregação, redução, novos atributos, etc e Identificar atributos úteis.



## Mineração

Algoritmos e técnicas de identificação de padrões nos dados e verificar hipóteses; Descritivas ou preditivas; Vários tipos diferentes.



## Interpretação

Análise dos resultados e tomada de decisão.





# Pré-Processamento



# Pré-Processamento

- Fase de tratamento dos dados de modo a anteceder o Machine Learning, para tal, é importante entender o contexto e a base de dados;
- Dados podem conter ruídos, imperfeições, inconsistências, duplicatas, ausências, com atributos independentes ou correlacionados, etc;
- Melhorar qualidade dos dados e eliminar quaisquer elementos que podem criar falsos resultados;

# Integração

- Dados podem ser oriundos de diversas fontes diferentes e em determinadas situações precisam ser integrados;
- Várias vezes esses dados será necessário a padronização desses dados;



# Eliminação Manual de Atributos

- Muitas vezes existem atributos que são irrelevantes para a análise, como alguns casos em que os nomes não são importantes (anonimização);
- Em análises preditivas, quando um atributo não contribui para a estimativa de um valor, ele é irrelevante para a análise e deve ser eliminado;
- Atributos que contém o mesmo valor para todos os objetos também devem ser eliminados, por exemplo, o campo cidade em uma base que analisa dados de uma determinada cidade;

# Amostragem de Dados

- Quanto mais dados, maior a precisão, menos a performance/eficiência;
- Haverá casos em que será necessário trabalhar apenas com uma amostra dos dados;
- Amostra Progressiva → Aumenta progressivamente até atingir um pico de precisão;

# Dados Desbalanceados

- É comum que dados de um subconjunto de uma determinada classe apareçam com frequência maior que as demais;
- Afeta muito o desempenho de alguns algoritmos de Machine Learning, favorecendo a classificação de novos dados na classe majoritária;

# Dados Desbalanceados

- Reduzir o tamanho do conjunto de dados, utilizando diferentes custos de classificação e induzir um modelo para uma classe;
- Algumas situações incluem as técnicas de classificação com apenas uma classe, ou os dados são treinados por classe (separadamente);

# Dados Incompletos

- Eliminar objetos com valores ausentes;
- Definir e preencher manualmente valores para atributos com valores ausentes;
- Usar métodos ou heurísticas;





# Dados Inconsistentes


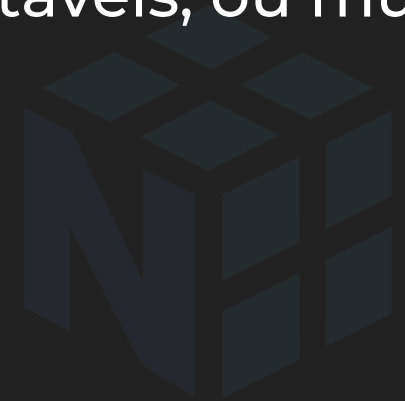
- Dados com valores conflitantes em seus atributos (Idade: 3; Peso: 120KG, por exemplo);
- Podem ser reconhecidas quando relações entre atributos são claramente conhecidas (valores correlacionados direta ou indiretamente);
- Algoritmos simples podem verificar a existência dos mesmos;
- Em casos de conjuntos não muito grandes, podem ser removidos manualmente;

# Dados Redundantes

- Objeto muito semelhante a outro no mesmo conjunto de dados;
- Também considerado redundante se puder ser deduzido a partir do valor de um ou mais atributos;
- Podem dar a falsa impressão de que esse perfil de objeto é mais importante que os outros, induzindo o modelo de análise ao erro;

# Dados com Ruídos



- Objetos que aparentemente não pertencem à distribuição que gerou os dados analisados;
  - Variância ou erro aleatório;
  - Outliers, valores acima dos limites aceitáveis, ou muito diferentes dos demais;
- 
- 

# Transformação de Dados

- Diversas técnicas de ML estão limitadas à valores em determinados tipos;
- Dependendo do modelo, alternar os valores entre qualitativos e quantitativos, nominais ou ordinais, etc;



# Transformação de Dados

- Redes Neurais Artificiais e Support Vector Machines, por exemplo, lidam apenas com dados numéricos;
- Limites inferior e superior de atributos são muito diferentes, ou atributos em escala diferente, há necessidade de transformação de valores numéricos para outro valor numérico, evita que um atributo predomine sobre outro;

# Redução da Dimensionalidade

- Dimensionalidade = tamanho horizontal do objeto;
- Em diversos algoritmos, grandes quantidades de atributos inviabilizam o processo;
- Melhora desempenho e torna resultados mais compreensíveis;





Vamos explorar na prática





OBRIGADO!

