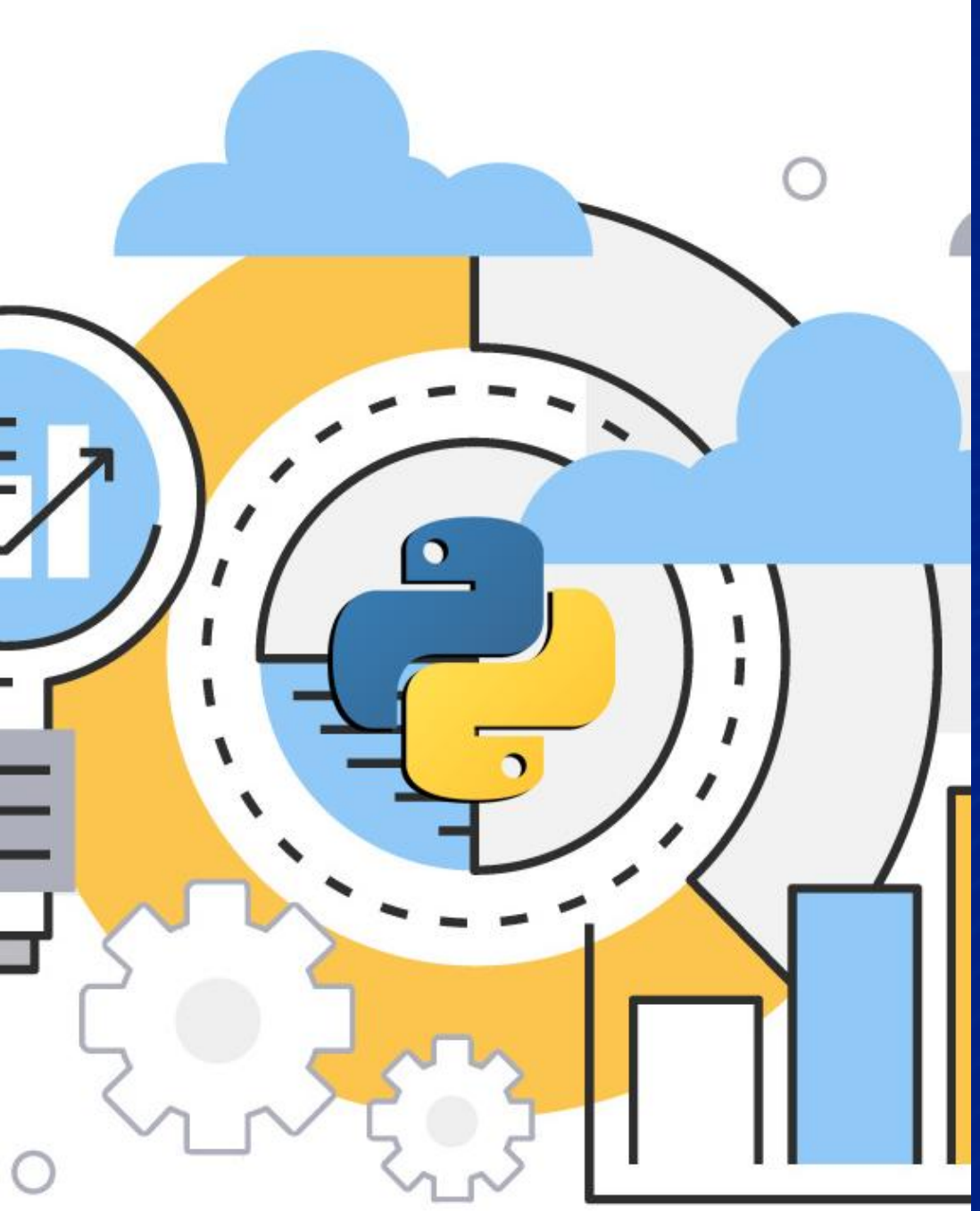
CIÊNCIA DE DADOS COM PYTHON

Prof. Renzo Paranaíba Mesquita

Inatel



CRITÉRIOS DE AVALIAÇÃO

- **C11**
  - A NP1 será formada pela prova PT1 com peso 70% e série de exercícios SE1 com peso 30%;
  - A NP2 será formada por um projeto prático PP1 com peso 35%, outro projeto prático PP2 com peso 35% e série de exercícios SE2 com peso 30%;
- NP3 – Prova com cobertura de todo conteúdo da disciplina, elaborada pelo professor.
- NÃO HAVERÁ SUB PARA EXERCÍCIOS E PROJETOS PRÁTICOS.

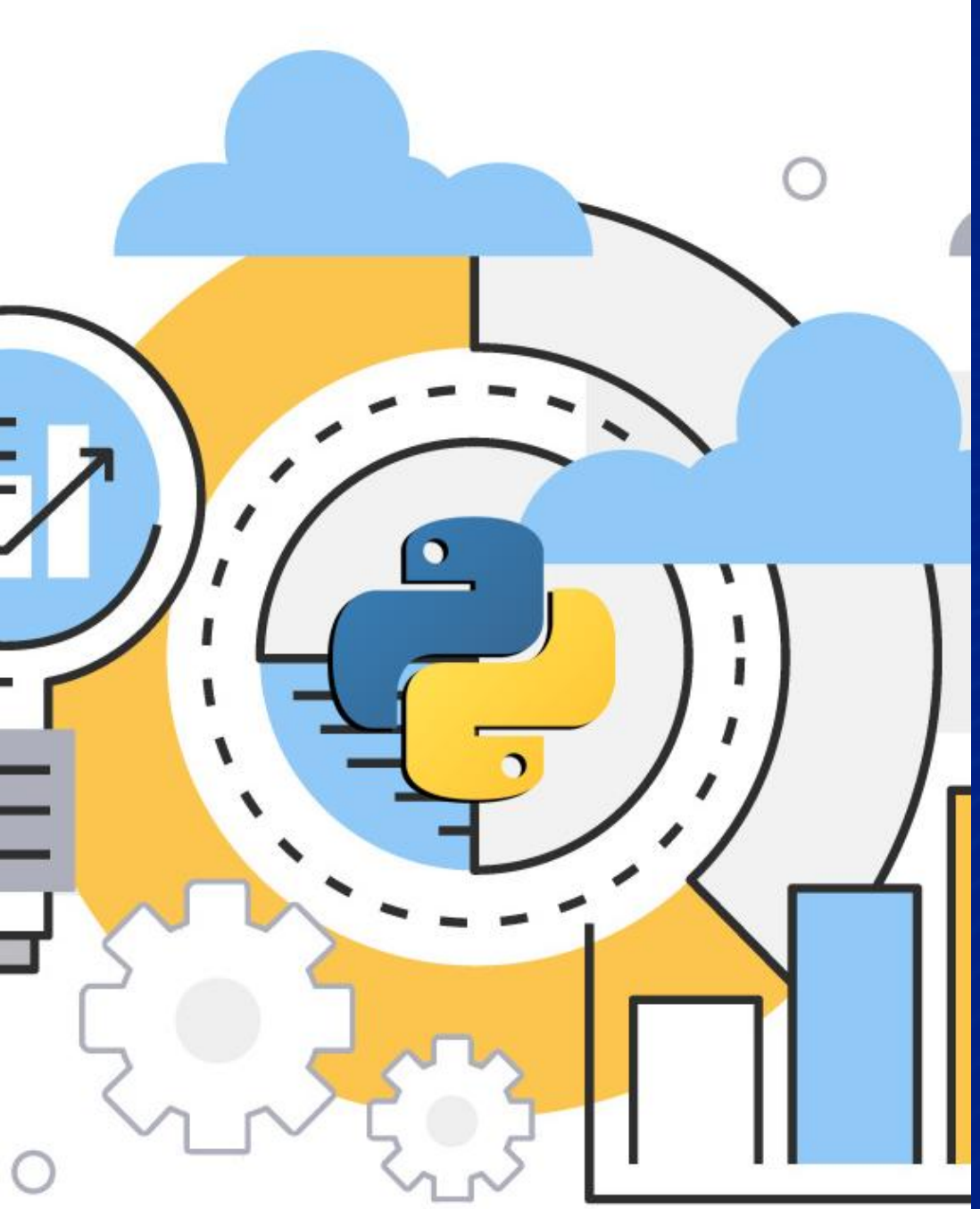
Inatel

CIÊNCIA DE DADOS COM PYTHON

CAP.1 - CONCEITOS E FERRAMENTAS FUNDAMENTAIS

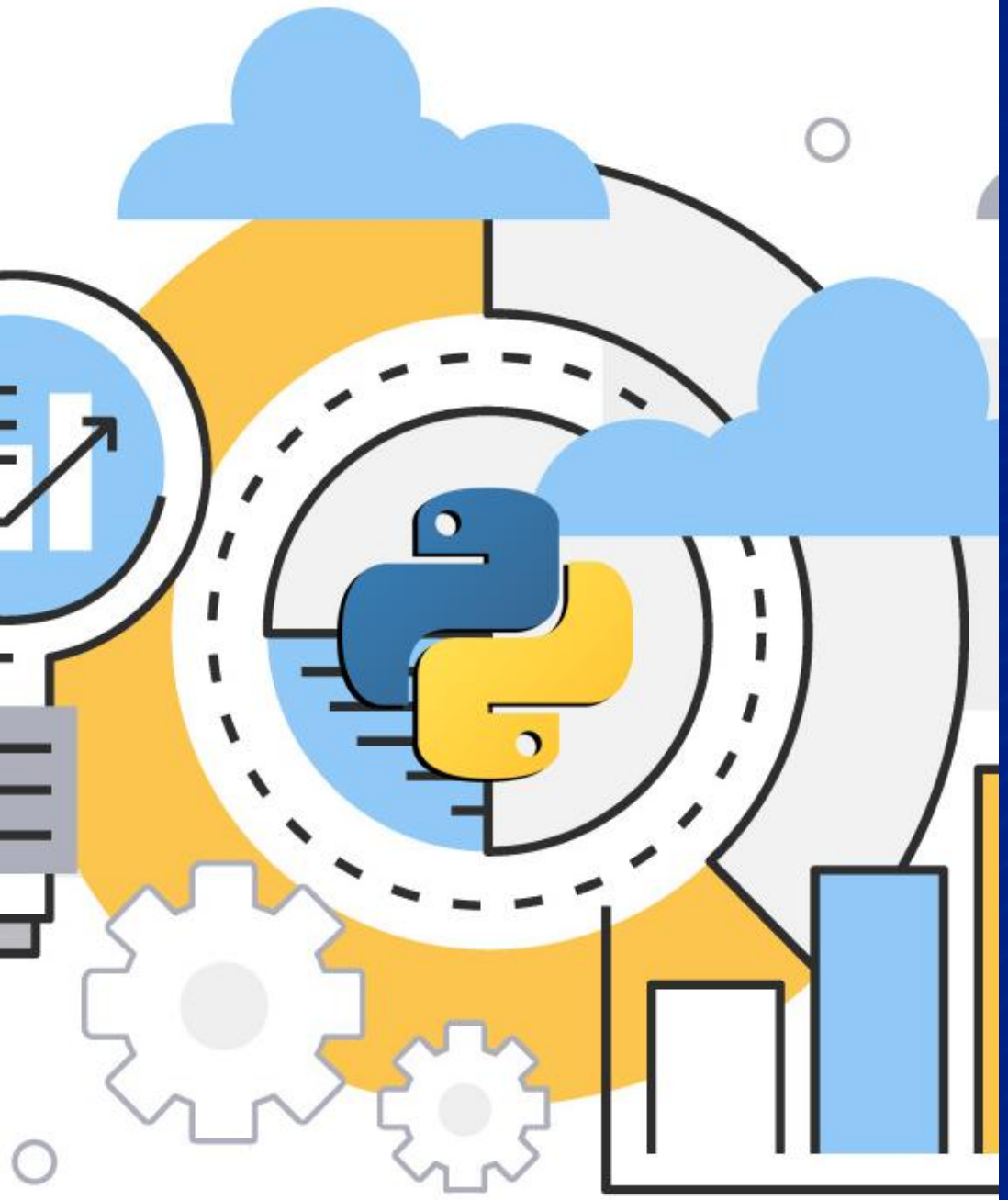
Prof. Renzo Paranaíba Mesquita

Inatel



OBJETIVOS

- Compreender o que é o BIG DATA;
- Entender o que é o campo da Ciência de Dados;
- Discutir as diferentes Variedades, Formatos e Fontes de Dados;
- Entender como a linguagem de Programação Python contribui para a Ciência de Dados;



TÓPICOS

1. O que é o BIG DATA?;
2. Conceito de Ciência de Dados (Data Science);
3. Casos de uso da Ciência de Dados;
4. Variedade de Dados (Data Varieties);
5. Análise de Dados (Data Analysis);
6. Formatos de Arquivos (Data Formats);
7. Fontes de Dados (Data Sources);
8. Por que usar Python?

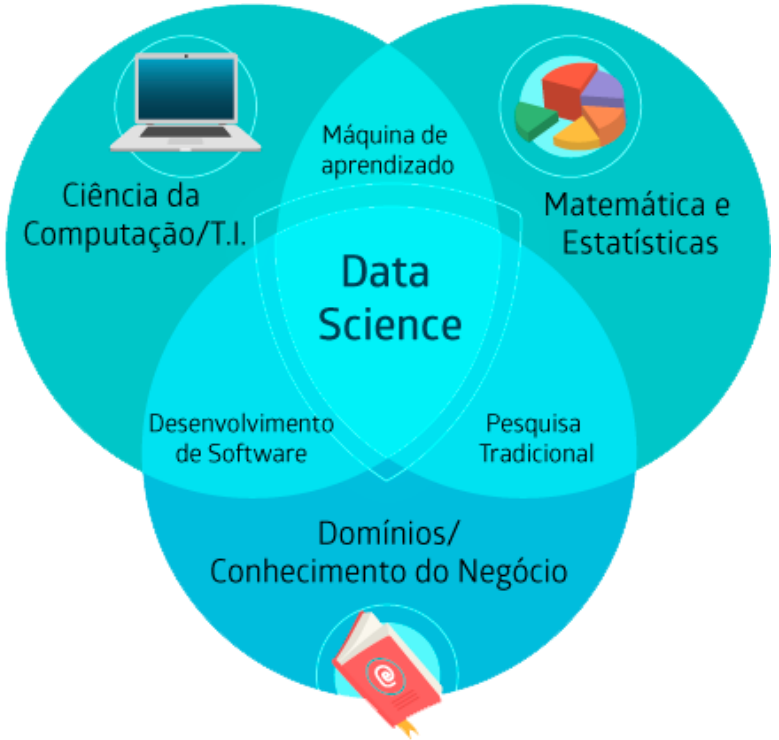
Inatel

1.1. O QUE É O BIG DATA?

- Neste exato momento, **o mundo está sendo inundado por novos dados** oriundos de:
  - Computadores pessoais;
  - Smartphones;
  - Câmeras;
  - Wearables;
  - Sensores;
  - Navegação na Internet, etc.
- TSUNÂMIS de dados estruturados, semiestruturados ou não estruturados **estão sendo produzidos por atividades que acontecem tanto no mundo real quanto no virtual**;
- Bem-vindo ao mundo do **BIG DATA!**

1.2. CONCEITO DE CIÊNCIA DE DADOS

- **Ciência de dados (CD) é um campo interdisciplinar** que combina Matemática, Estatística, Ciência da computação e Conhecimento Especializado para extrair *insights* e conhecimento significativos dos dados;
- A CD envolve processos como **coleta de dados, limpeza, exploração, modelagem e visualização** para analisar tendências, fazer previsões e orientar a tomada de decisões de diferentes negócios;
- A área deu até mesmo origem a **profissão de Cientista de Dados**, ou seja, profissionais capazes de extraírem *insights* valiosos dos dados para resolver problemas de diferentes naturezas;



1.3. CASOS DE USO DA CIÊNCIA DE DADOS

"Bradesco cria sistema antifraudes analisando *logs* gerados por sensores em caixas eletrônicos". Como resultado, conseguiu reduzir de 10 mil para 5 o número de incidentes diários.

"Airbnb se torna maior empresa hoteleira da atualidade, mas sem possuir nenhum hotel". Grande parte disto se deve à utilização de um modelo orientado a dados (*data-driven)* para tomada de decisões estratégicas.

"Nike, gigante dos artigos esportivos, adquiriu a empresa Celect" para reunir e tratar dados de seus clientes, objetivando identificar tendências e adaptar seus produtos de acordo com as demandas do mercado.

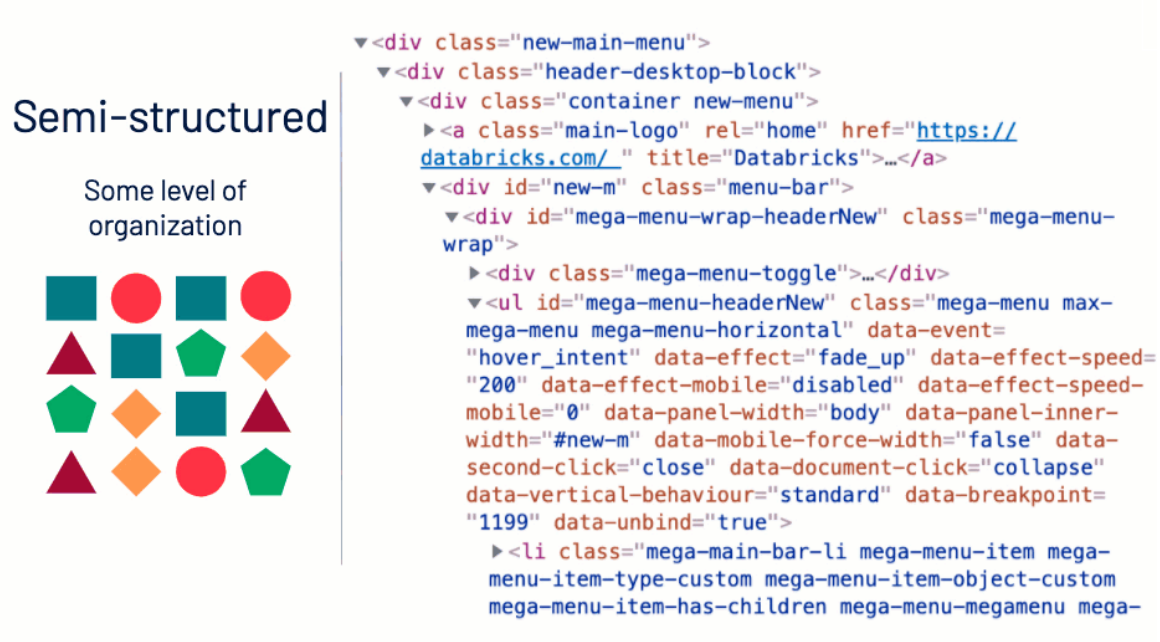
1.4. VARIEDADE DE DADOS (DATA VARIETIES)

- O ideal seria que todos os dados a serem analisados já estivessem em repositórios organizados, mas...
- Dados podem ser de três tipos: **Estruturados, Semiestruturados ou Não Estruturados;**
- **Dados Estruturados (Structured Data):**
  - Organizados e representados em uma estrutura previamente planejada (**Schema**);
  - **Bancos de Dados Relacionais** são exemplos clássicos de dados estruturados;

Order	CustID	Month	Item	Color	Price
101	20051	Dec	Pen	Red	2.99
102	20045	Mar	Pencil	Blue|Yellow|Red	3.99
103	29584	May	Eraser	Blue	1.25
104	29584	May	Pen	White	2.25
105	29584	May	Pencil	Blue|Yellow|Red	2.99
106	27485	Jan	Eraser	Blue|Yellow	2.75
107	29574	Jan	Marker	Green	1.75
108	24447	Feb	Marker	Yellow|Blue	7.25
109	26466	Jul	Pen	Black|Red	5.25
110	27467	Jun	Pencil	Black	2.95

1.4. VARIEDADE DE DADOS (DATA VARIETIES)

- O ideal seria que todos os dados a serem analisados já estivessem em repositórios organizados, mas...
- Dados podem ser de três tipos: **Estruturados, Semiestruturados e Não Estruturados;**
- **Dados Semiestruturados (Semi-structured Data):**
  - Não possuem uma estrutura pré-planejada, mas **contém *tags* ou outros tipos de marcadores** para separar e identificar elementos;
  - Arquivos **XML, JSON e CSV** são exemplos clássicos de dados semiestruturados.



1.4. VARIEDADE DE DADOS (DATA VARIETIES)

- O ideal seria que todos os dados a serem analisados já estivessem em repositórios organizados, mas...
- Dados podem ser de três tipos: **Estruturados, Semiestruturados e Não Estruturados;**
- **Dados Não Estruturados (Unstructured Data):**
  - São dados sem padrão ou estrutura;
  - Arquivos como documentos, imagens, fotos e vídeos são exemplos clássicos de dados não estruturados;

1.5. ANÁLISE DE DADOS

- A Análise de Dados é uma das etapas mais importantes da Ciência de Dados e foco deste nosso curso;
- Geralmente, organizada nas seguintes etapas:

Conscientização	Compreensão	Descoberta	Análise	Divulgação
1. Que perguntas eu preciso responder para ter sucesso na minha pesquisa ou negócio? (Conscientização)	2. Quais *datasets* poderiam me ajudar a responder estas perguntas e quão úteis eles são? (Compreensão)	3. Limpe, Explore e misture os *datasets* por meio de técnicas e ferramentas poderosas para dar suporte à análise; (Descoberta)	4. Compreenda o relacionamento entre os dados em diferentes dimensões e plote as descobertas de maneira amigável; (Análise)	5. Divulgue os resultados aos interessados de uma forma clara e objetiva a fim de responder as perguntas levantadas e produzir *insights* (Divulgação)

Inatel

1.6. FORMATOS DE ARQUIVOS (FILE FORMATS)

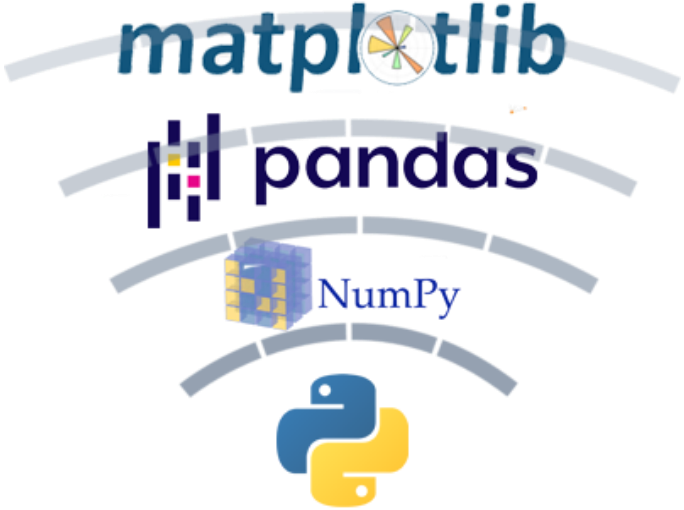
- São inúmeros os formatos disponíveis para extração de dados valiosos;
- Alguns formatos populares: XLSX, DOCX, CSV, JSON, XML e SQL.

1.7. FONTES DE DADOS (DATA SOURCES)

- Existem conjuntos de dados de todos os tipos, sejam públicos ou privados, mais gerais ou específicos;
- Alguns exemplos de fontes de dados públicas para buscarmos *Datasets*:
  - Kaggle;
  - Google Dataset Search;
  - dados.gov.br (dados do governo brasileiro);
  - data.gov (dados do governo americano).

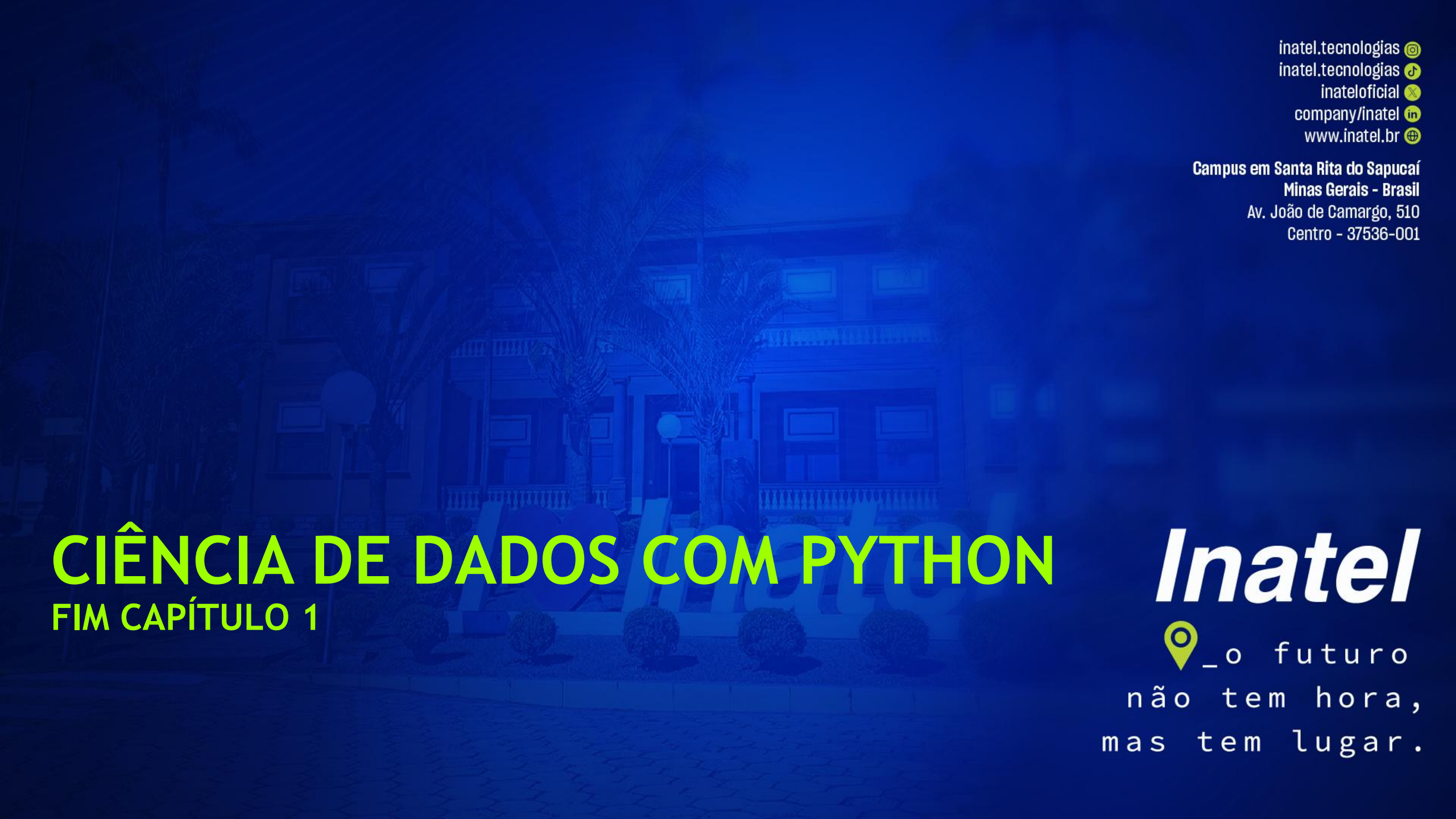
1.8. POR QUE USAR PYTHON?

- Python é a principal linguagem de programação voltada para Ciência de Dados;
- O grande diferencial da linguagem se encontra nas bibliotecas otimizadas que ela oferece para Análise de Dados;
- Dentre muitas, destacam-se as seguintes e foco deste curso:
  - **NumPy**
    - Biblioteca fundamental para se trabalhar com Arrays Multidimensionais e que oferece um grande conjunto funções matemáticas;
  - **Pandas**
    - Oferece estruturas e funções poderosas para manipular tabelas de dados e séries temporais;
  - **Matplotlib**
    - Principal biblioteca do Python para plotagem de gráficos.



1.8. POR QUE USAR PYTHON?

- Sugestões de Ambientes populares para se trabalhar com Análise de Dados com Python:
- Ambiente Offline: PyCharm Community Edition
  - https://www.jetbrains.com/pycharm/download/?section=windows
- Ambiente Online: Google Colab
  - https://colab.google/



inatel.tecnologias
inatel.tecnologias
inateloficial
company/inatel
www.inatel.br

Campus em Santa Rita do Sapucaí
Minas Gerais - Brasil
Av. João de Camargo, 510
Centro - 37536-001

CIÊNCIA DE DADOS COM PYTHON

FIM CAPÍTULO 1

Inatel

_o futuro
não tem hora,
mas tem lugar.