# Introdução à Ciência de Dados em Cibersegurança

**Ligia F. Borges e Michele Nogueira**

[1] Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG)

{ligia.borges,michele}@dcc.ufmg.br

***Resumo.*** *A evolução e a popularização dos dispositivos computacionais e tecnologias de comunicação permitem o surgimento de novas ameaças de segurança ou potencializam as ameaças existentes. Pessoas mal-intencionadas exploram esses avanços e o aperfeiçoamento de técnicas de análise de dados e a inteligência artificial para criar uma quantidade maior de ataques, muitos deles com maior sofisticação, maior frequência e abrangência. Esses ataques ameaçam a integridade, a disponibilidade e a privacidade dos dados e dos sistemas digitais. Assim, a disseminação do tema e o treinamento de profissionais considerando os avanços e perspectivas da área são cada vez mais necessários. Este tutorial contribui nesta direção atraindo, em um primeiro momento, a atenção ao tema de ciência de dados aplicada à cibersegurança. O tutorial revisa as principais definições e fundamentos sobre cibersegurança, ciência de dados e fundamentos de estatística. Então, estudos de caso relevantes são apresentados e discutidos.*

***Abstract.*** *The evolution and popularization of computational devices and communication technologies allow for the emergence of new security threats or amplify existing ones. Malicious individuals exploit these advancements and the refinement of data analysis and artificial intelligence techniques to create a greater quantity of attacks, many of them with higher sophistication, frequency, and scope. These attacks jeopardize the integrity, availability, and privacy of data and digital systems. Therefore, the dissemination of the topic and the training of professionals, considering the advancements and perspectives of the field, are increasingly necessary. This tutorial contributes in this direction, initially drawing attention to the theme of data science applied to cyber security. The tutorial reviews the main definitions and foundations of cyber security, data science, and statistics. Then, relevant case studies are presented and discussed.*

## 1. Identificação do Tutorial

1. **Título:** Introdução à Ciência de Dados em Cibersegurança
2. **Autores:**
    - **Dra. Michele Nogueira Lima** (apresentadora)
      Universidade Federal de Minas Gerais
      michele@dcc.ufmg.br
    - **Dra. Ligia F. Borges** (apresentadora)
      Universidade Federal de Minas Gerais

ligia.borges@dcc.ufmg.br

3. **Tipo:** O tutorial será prioritariamente teórico, porém serão apresentados exemplos práticos usando a plataforma online Kaggle. Os participantes serão convidados a criar uma conta no Kaggle e para um melhor aproveitamento sugere-se aos participantes que tragam seus laptops com a possibilidade de acesso à rede sem fio para acompanhamento dos exemplos.
4. **Idioma da apresentação:** Português.
5. **Equipamento requerido:** Projetor multimídia, quadro branco, pincel e acesso à Internet sem fio essencial para os apresentadores e desejável para os participantes.

## 2. Visão geral do tutorial

A cibersegurança é um dos pilares fundamentais da era digital, especialmente com a crescente dependência na tecnologia e na Internet em nossas vidas. A popularização dos dispositivos computacionais através da Internet das Coisas e de tecnologias de comunicação permite a oferta diária de diversos serviços disponibilizados pelas redes, gerando mais e mais dados assim como uma maior dependência das pessoas e empresas nas tecnologias digitais. Estima-se que nos últimos dois anos uma taxa de 90% dos dados existentes foi gerada. Isso inclui desde dados sensoriados continuamente pelos diversos dispositivos computacionais em rede, como câmeras de monitoramento, sensores, até fotos e vídeos em mídias sociais, carrinhos de compras, entre muitos outros. Com mais dados à disposição e uma maior dependência nos serviços digitais, surge o grande desafio de garantir a segurança, a integridade e a privacidade dos dados e dos sistemas.

À medida que navegamos pela era digital, a interseção entre ciência de dados, incluindo as técnicas de inteligência artificial (IA) e aprendizado de máquina, e a cibersegurança tornou-se um tópico de suma importância. O grande volume de dados disponíveis e as melhorias significativas no poder computacional e nos métodos de aprendizado de máquina fazem com que a IA impacte de forma transformadora a cibersegurança. Cabe aqui mencionar que o uso da ciência de dados possui duas perspectivas no contexto de segurança cibernética: 1) sua aplicação para construir sistemas defensivos, como detecção de *malware* e ataques de rede; e 2) sua exploração para lançar ataques mais velozes, volumosos, com maior abrangência e sofisticação. Este tutorial foca na perspectiva 1, além de também enfatizar a necessidade de construirmos sistemas de IA robustos e imunes à interferência externa.

A ciência de dados, particularmente a IA e o aprendizado de máquina, vem transformando a segurança cibernética. Ela oferece recursos para analisar grandes quantidades de dados, identificar padrões e fazer previsões, tudo com uma velocidade e precisão que superam as capacidades humanas. Essa capacidade é crucial para detectar e mitigar ameaças cibernéticas em tempo real, melhorando assim a postura geral de segurança de uma organização. Por exemplo, a IA ajuda a identificar padrões anômalos indicativos de um ataque cibernético em seus primórdios. Isso é particularmente importante devido ao surgimento de métodos de ataque sofisticados, como *malware* sem arquivo, e os desafios de segurança impostos por plataformas de computação. Os modelos de segurança tradicionais, que dependem da verificação de hashes de arquivo em relação a amostras de *malware* conhecidas, não são suficientes para combater essas ameaças avançadas.

A academia já possui um longo histórico de atuação na integração da inteligência artificial e a segurança cibernética. Por anos, várias técnicas de IA, como redes neurais, algoritmos genéticos e aprendizado de máquina, são aplicadas em sistemas de detecção de intrusão, detecção de anomalias em tráfego de rede e na classificação de emails e identificação de SPAMs. Atualmente, várias empresas se beneficiam da IA para reforçar suas defesas de segurança cibernética. Por exemplo, a plataforma Reveal(x) da ExtraHop realiza uma análise baseada em regras e comportamentos para fornecer informações sobre o tráfego de rede e identificar possíveis ameaças. Da mesma forma, o Vectra Cognito usa IA para detectar ameaças futuras ou desconhecidas com base em uma análise de cargas de trabalho e técnicas de *malware* conhecidas.

Embora a ciência de dados apresente um grande potencial para a cibersegurança, ela não é uma bala de prata. Ela deve ser vista como uma ferramenta que complementa, e não substitui, as medidas de segurança tradicionais. Além disso, o seu uso efetivo na segurança cibernética requer uma compreensão profunda da tecnologia, das técncias e do cenário de ameaças cibernéticas. À medida que avançamos, podemos esperar ver mais integração da ciência de dados com outras tecnologias. É provável que os sistemas de gerenciamento de informações e eventos de segurança (SIEM) integrem os dados do usuário para fornecer uma visão mais abrangente dos eventos de segurança. Ao aproveitar o potencial da ciência de dados, as organizações aprimoraram sua capacidade de detectar e responder a ameaças cibernéticas, protegendo assim os sistemas diante de um cenário cibernético cada vez mais complexo. Além disso, é importante ressaltar que a interseção da ciência de dados e a cibersegurança apresenta um terreno fértil para inovações, mas também traz novos desafios, como a possível existência de viés nos dados, a explicabilidade, os ataques adversariais e a privacidade.

Neste tutorial, exploramos como a IA e técnicas de aprendizado de máquina estão sendo aplicadas para proteger os sistemas contra ameaças cibernéticas. Apresentamos como a ciência de dados tem automatizado processos e aprimorado a detecção e resposta a ameaças, passando pela análise do comportamento de usuários e sistemas para identificar atividades suspeitas ou anomalias indicativas de ataques e identificação de vulnerabilidades. O tutorial inicia com uma parte teórica sobre cibersegurança e ciência de dados destacando seu uso na prevenção, detecção e mitigação de ataques. Será apresentada a plataforma Kaggle, utilizada em ciência de dados, através de exemplos na área de cibersegurança. O tutorial segue com uma série de diferentes estudos de caso, em Python, representativos da área de cibersegurança e relacionados aos principais problemas, como a detecção de *malware*, detecção de anomalia, de bots, de spams, vazamentos de informações e *phishing*. O tutorial segue uma abordagem introdutória e não se aprofundará nos tópicos abordados, porém oferecerá uma visão geral dos avanços e conceitos relacionados ao tema.

### 2.1. Justificativa

Diante das mudanças progressivas em cibersegurança, muitas delas impulsionadas pela ciência de dados e pela oferta de mais recursos computacionais, é cada vez mais necessária a disseminação do tema e a formação e/ou atualização de profissionais que conhecem as vantagens, desafios e as limitações do uso de inteligência artificial e outras técnicas de ciência de dados na área de cibersegurança. A importância do treinamento desses profissionais é ressaltada inclusive na edição de maio de 2023 do relatório anual *'The Future*

*of Jobs'*, divulgada pelo Fórum Econômico Mundial, em que indica para uma escala de cinco anos as 10 principais profissões do futuro, dentre elas encontram-se analista de segurança da informação, analista e cientista de dados e especialista em inteligência artificial e aprendizado de máquina. Este tutorial tem como objetivo oferecer aos participantes um primeiro contato com o tema de ciência de dados aplicada à cibersegurança, fornecendo uma visão geral dos avanços e conceitos relacionados ao tema.

### 2.2. Perfil desejado do público alvo

Este tutorial destina-se a estudantes de Ciência da Computação ou Engenharia da Computação, pesquisadores e profissionais graduados da área de redes de computadores, telecomunicações ou afins interessados em conhecer os benefícios e os desafios relacionados à promissora evolução que vem ocorrendo em ciência de dados e seu uso em cibersegurança. É aconselhável que os participantes possuam conhecimentos básicos, em nível de graduação, em estatística, redes de computadores e programação.

### 2.3. Tutoriais recentes relacionados

Dentre os tutoriais e minicursos recentes relacionados a este tutorial estão o minicurso intitulado "*Das Redes Vestíveis aos Sistemas Ciber-Humanos: Uma Perspectiva na Comunicação e Privacidade dos Dados"* apresentado no Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC 2021) fornece uma visão geral da evolução e recentes avanços dos sistemas ciberfísicos e ciber-humanos. O trabalho abordou questões relacionadas as vulnerabilidades de segurança encontradas nesses sistemas, os problemas relacionados a privacidade dos dados e garantia de disponibilidade dos serviços, bem como as ameaças encontradas. Além desse minicurso, as palestras intituladas *Empowering DDoS Attack Prediction through Machine Learning and Deep Learning*, *Data Science for Cybersecurity: An Overview Focused on Networking* e *Security, Privacy and Resilience in the Internet of Health Things*, relacionadas a este tutorial, têm sido proferidas em diferentes instituições internacionais e nacionais e eventos científicos.

## 3. Escopo e estrutura

O conteúdo do tutorial será introduzido de forma aplicada e através de uma perspectiva prática. O tutorial se divide em três partes, iniciando pela introdução de conceitos e definições relevantes relacionados à cibersegurança, ciência de dados e estatística. Na sequência, as principais ferramentas serão introduzidas, incluindo noções básicas de programação em Python e as principais bibliotecas. Por fim, o tutorial finaliza com um conjunto de estudos de caso específicos da área de cibersegurança em que os conhecimentos adquiridos serão combinados e utilizados na análise dos dados, predição e detecção de ataques. O conteúdo detalhado de cada parte segue:

- **Parte 1 - Introdução:** definições e fundamentos sobre cibersegurança (princípios de segurança, principais ataques e defesas), ciência de dados (incluindo conceitos referentes à aprendizado de máquina supervisionado, não supervisionado e semi-supervisionado e aprendizado estatístico) e fundamentos de estatística (medidas de tendência central e medidas de dispersão), curva ROC, matriz de confusão e Kappa, acurácia.
- **Parte 2 - Ferramentas:** básico sobre plataformas Kaggle e Jupyter e uso de bibliotecas como pandas e numpy. A introdução às ferramentas ocorrerá por meio de exemplos na área de cibersegurança.

- **Parte 3 - Estudos de caso:** Este módulo é composto por quatro estudos de caso, cada estudo segue uma parte teórica e outra parte prática. Os cinco estudos são: $(i)$ introdução a malware e aplicação de aprendizado de máquina para detecção de malwares; $(ii)$ sistemas de detecção de intrusão e exemplo prático do uso de técnicas de aprendizado na detecção de intrusão; $(iii)$ ataques de negação de serviço e predição de ataques através de aprendizado estatístico; $(iv)$ SPAMs e técnicas de classificação na detecção de SPAMs.

## 4. Bibliografia principal


1. Sikos, L.F. and Raymond-Choo, K.K., "Data Science in Cybersecurity and Cyberthreat Intelligence". Intelligent Systems Reference Library, Springer, 2020.
2. James, G., Witten, D., Hastie, T., Tibshirani R., "An Introduction to Statistical Learning". Second edition.
3. Heard, N., Adams, N., Rubin-Delanchy, P., Turcotte, M., "Data Science for Cyber-Security". 2019 World Scientific Publishing Europe Ltd.
4. Brooks, C. J., Grow, C., Craig, P., Short, D., "Cybersecurity Essentials", 2018, John Wiley & Sons, Inc.
5. Cunningham, C. and Touhill, G. J., "Cyber Warfare - Truth, Tactics, and Strategies: Strategic concepts and truths to help you and your organization survive on the battleground of cyber warfare", 2020, Packt Publishing.
6. Pelloso, M., Vergutz, A., Santos, A. and Nogueira, M., "A Self-Adaptable System for DDoS Attack Prediction Based on the Metastability Theory", IEEE Global Communications Conference, Abu Dhabi, United Arab Emirates, 2018, pp. 1-6.
7. Nogueira, M., Santos, A., and Moura, J.M.F., "Early Signals from Volumetric DDoS Attacks: An Empirical Study", eprint 1609.09560, arXiv. 2016.
8. Sarker, I.H., Kayes, A.S.M., Badsha, S., "Cybersecurity data science: an overview from machine learning perspective". J. Big Data 7, 41 (2020). `https://doi.org/10.1186/s40537-020-00318-5`
9. Mvula, P.K., Branco, P., Jourdan, G.V., "A systematic literature review of cyber-security data repositories and performance assessment metrics for semi-supervised learning". Discover Data 1, 4 (2023). `https://doi.org/10.1007/s44248-023-00003-x`

## 5. Currículo das autoras

**Michele Nogueira Lima - `http://lattes.cnpq.br/2934786440085983`**

Pesquisadora de Produtividade em Pesquisa do CNPq, 1D, atua nas áreas de redes de computadores e segurança de redes. Possui doutorado em Ciência da Computação pela Sorbonne Université - UPMC/LIP6, França e realizou Pós-doutorado na Universidade Carnegie Mellon (CMU), EUA. É professora associada do Departamento de Ciência da Computação da Universidade Federal de Minas Gerais (UFMG). Suas pesquisas têm resultado na proposição de modelos matemáticos, protocolos e arquiteturas de sistemas. Foi membro do comitê consultivo da Comissão Especial em Segurança da Informação e de Sistemas Computacionais da Sociedade Brasileira de Computação. Foi editora técnica associada do periódico "IEEE Communications Magazine"(2012-2020), e é editora associada dos periódicos "Computer Communications", "Journal of Network and Systems Management"e "IEEE Communications Surveys & Tutorials". Coordenou o Comitê Técnico da Internet do IEEE ComSoc e a Comissão Especial de Segurança da Informação e de Sistemas Computacionais (CESeg) da SBC. Tem vasta experiência em proferir palestras, tutoriais e minicursos no Brasil e no exterior. É uma dos palestrantes distintos do *IEEE Communications Society*.

**Ligia Francielle Borges - `http://lattes.cnpq.br/0195128157754342`**

Pós-doutoranda no Departamento de Ciência da Computação da UFMG. Doutora pelo Programa de Pós-Graduação em Informática da Universidade Federal do Paraná (UFPR) e integrante do Centro de Ciência de Segurança Computacional (CCSC). Mestre em Tecnologias Computacionais para o Agronegócio pela Universidade Tecnológica Federal do Paraná – Câmpus Medianeira, 2017 e Especialista em Redes de Computadores: Projeto e Implementação pela Universidade Tecnológica Federal do Paraná – Câmpus Cornélio Procópio, 2014. Tecnóloga em Redes de Computadores pelo Centro de Ensino Superior de Foz do Iguaçu, 2013.