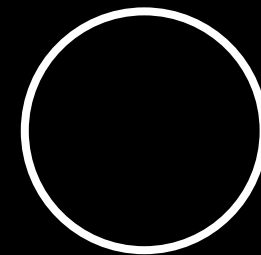




PROTECTING SENSITIVE INFORMATION



AGENDA



- ❖ Best Practices on how to protect company and personal data
- ❖ Protection techniques throughout the data pipeline
- ❖ Deep Dive into preprocessing & anonymization
- ❖ Machine Learning models gone rogue



AGENDA



- ❖ Best Practices on how to protect company and personal data
- ❖ Protection techniques throughout the data pipeline
- ❖ Deep Dive into preprocessing & anonymization
- ❖ Machine Learning models gone rogue



Panel Discussion with Marie Jadhvani and Pete Stiglich



INTRODUCTION



Monica Kay Royal

Founder & Chief Data Enthusiast



nerdnourishment

- ★ 16+ years experience
- ★ Accounting & Management Information Systems
- ★ Security/Risk/Compliance
- ★ CPA, CISA, CISM, AAIA
- ★ Data Career Coaching
- ★ Instructor at LinkedIn Learning
- ★ Podcast Host: Data Podcast for Nerds!
- ★ Board Member with ISACA: Academic Relations



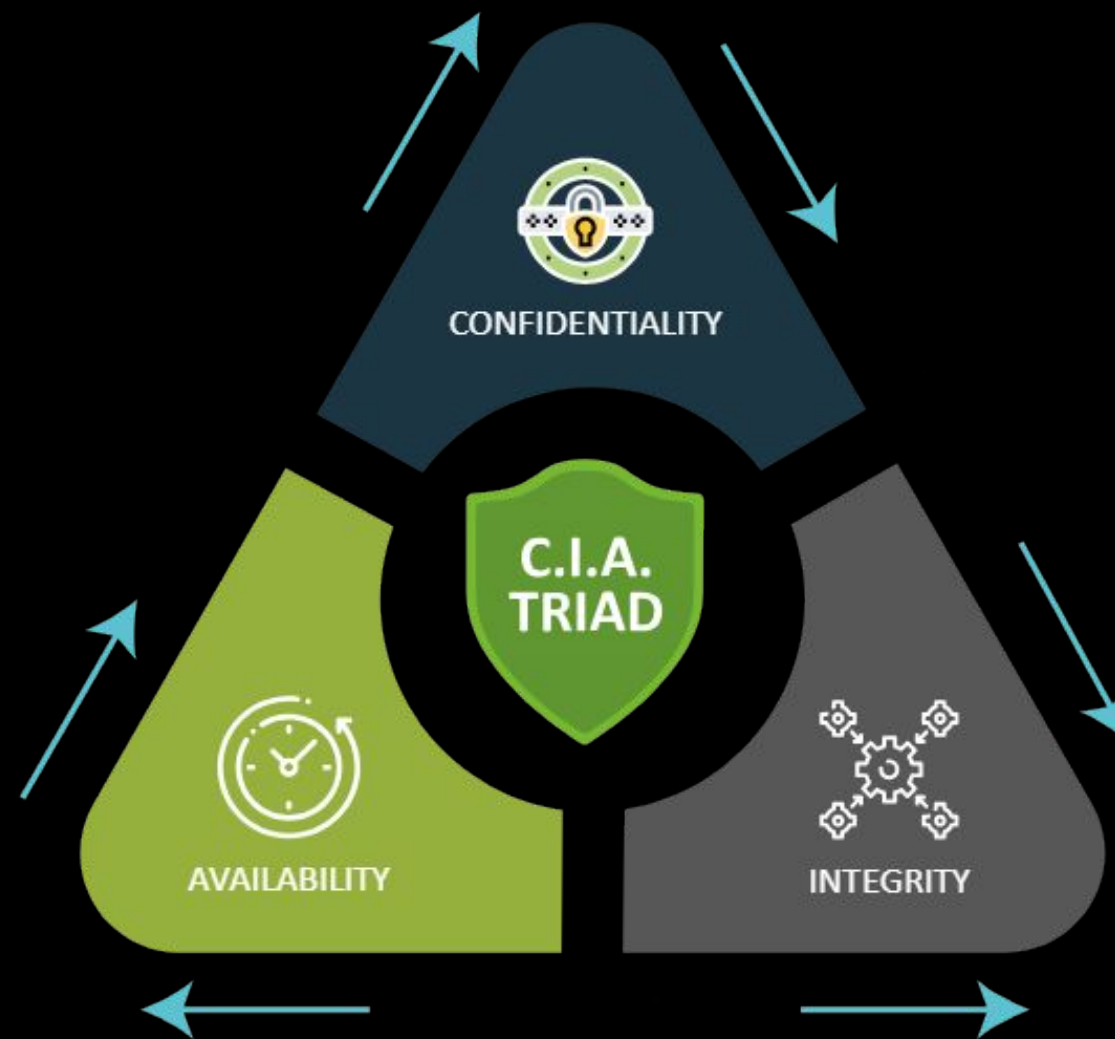
WHAT IS DATA SECURITY



WHAT IS DATA SECURITY

Process of maintaining the confidentiality, integrity, and availability of an organization's data

WHAT IS DATA SECURITY

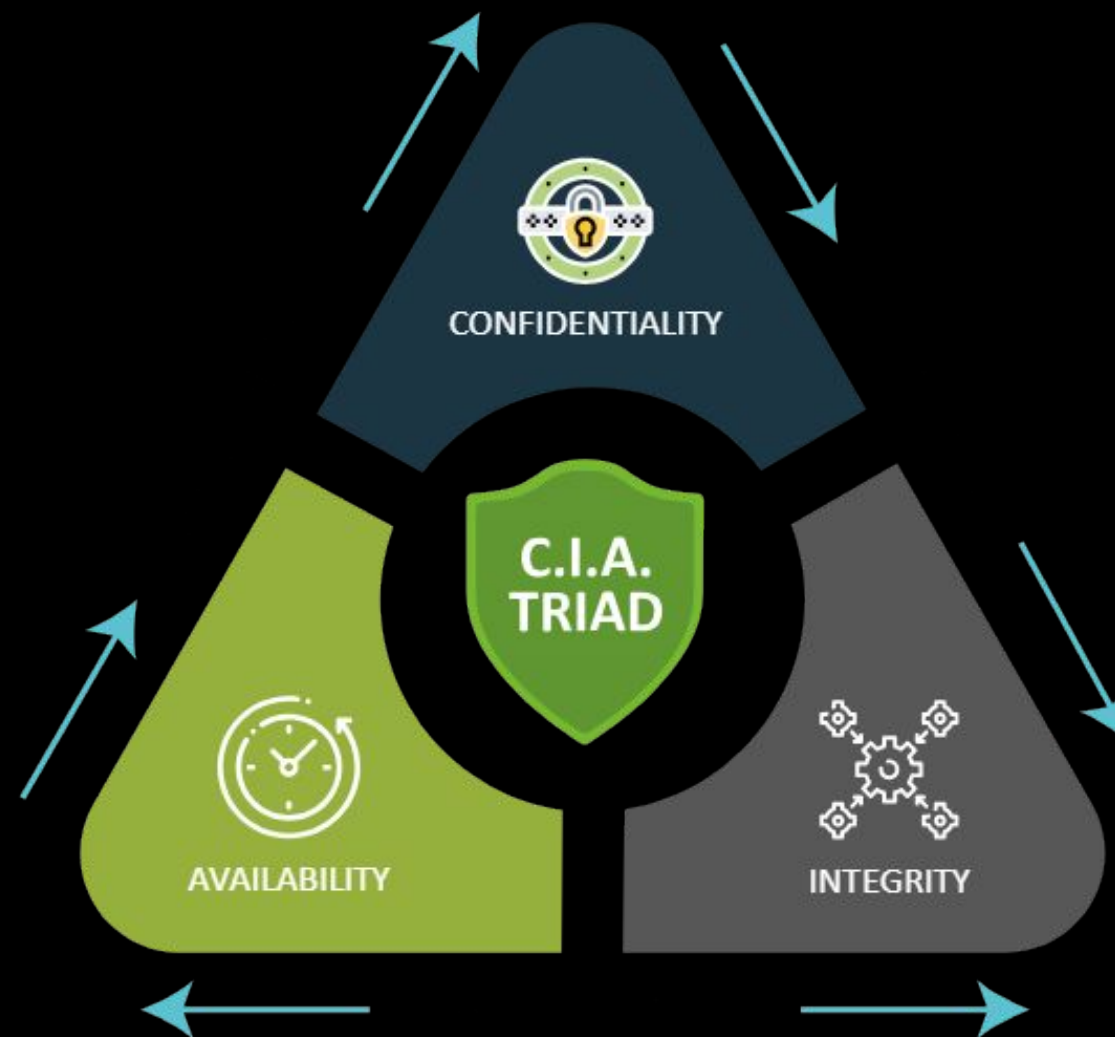


<https://websitecuritystore.com/wp-content/uploads/2021/08/cia-triad.svg>

WHAT IS DATA SECURITY

a.k.a Privacy

Preventing sensitive information from unauthorized access attempts

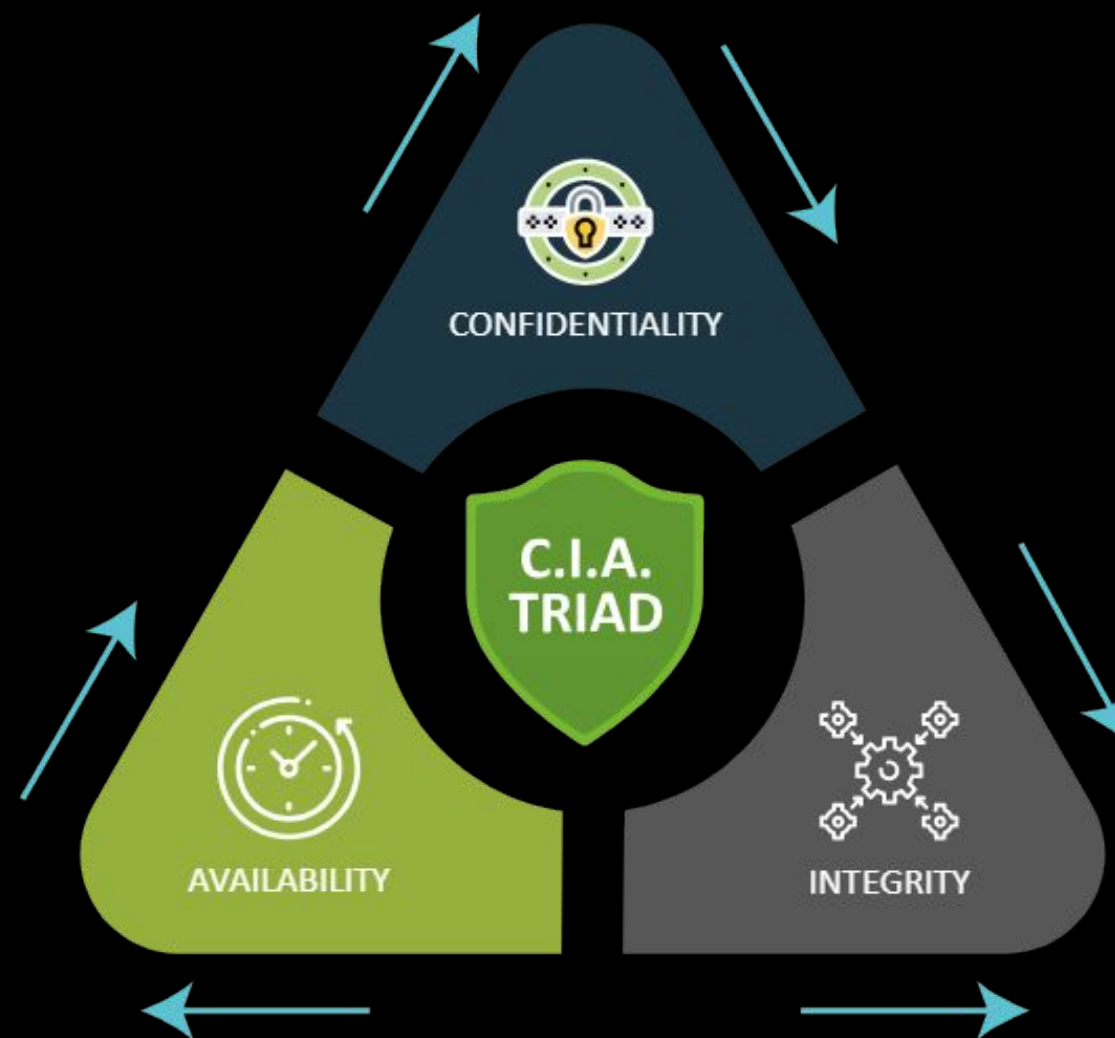


<https://websitecuritystore.com/wp-content/uploads/2021/08/cia-triad.svg>

WHAT IS DATA SECURITY

a.k.a Privacy

Preventing sensitive information from unauthorized access attempts



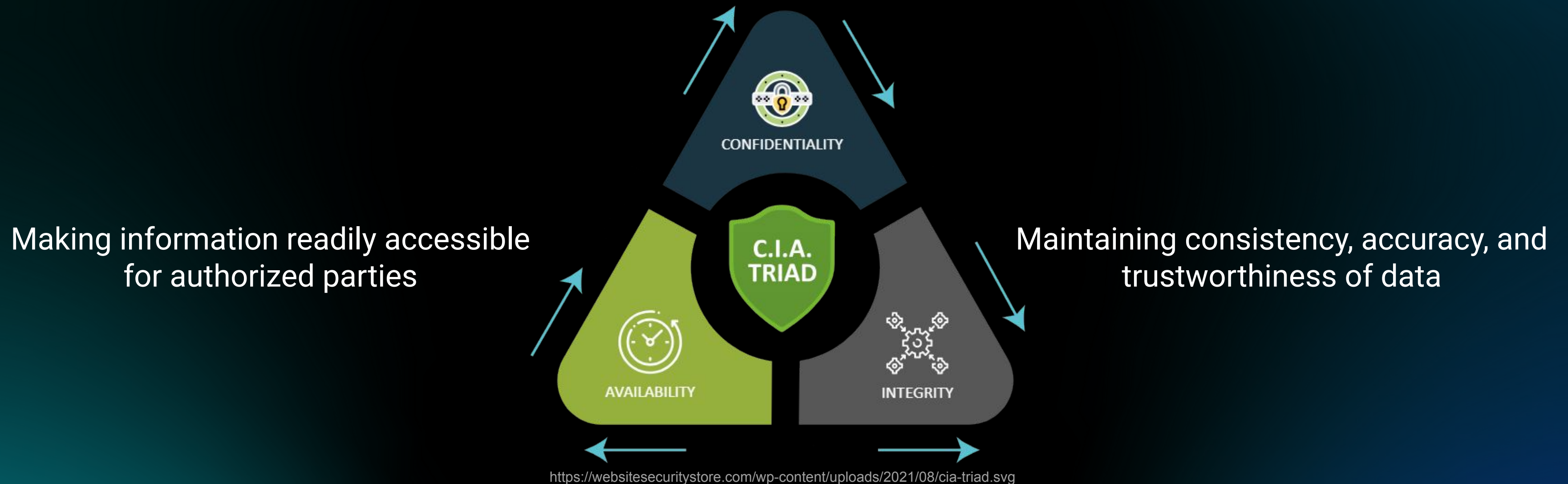
Maintaining consistency, accuracy, and trustworthiness of data

<https://websitecuritystore.com/wp-content/uploads/2021/08/cia-triad.svg>

WHAT IS DATA SECURITY

a.k.a Privacy

Preventing sensitive information from unauthorized access attempts



Information security is the practice of protecting information, with a **goal** to ensure the safety and privacy of critical and sensitive data.



INFORMATION **SECURITY**

Great Data Great Responsibilities!



https://media.licdn.com/dms/image/D4E12AQOXNk7yMAI_JQ/article-cover_image-shrink_600_2000/0/1679530882012?e=2147483647&v=beta&t=KLAspc8LkSiErIpZ7kfcU6-C-a01k_ThV8phlzaaUMI

Great Data Great Responsibilities!



Protecting Sensitive Information

DATA ANALYTICS LIFECYCLE

DATA ANALYTICS **LIFECYCLE**



DATA ANALYTICS **LIFECYCLE**



need to protect

**Data
Access**

**Data
Collection**

**Data
Preprocessing**

**Data
Sharing**

**Data
Storage**

**Data
Disposal**

DATA ANALYTICS **LIFECYCLE**



need to protect

**Data
Access**

**Data
Collection**

**Data
Preprocessing**

**Data
Sharing**

**Data
Storage**

**Data
Disposal**



DATA ANALYTICS **LIFECYCLE**



need to protect

**Data
Access**



**Data
Collection**



**Data
Preprocessing**

**Data
Sharing**

**Data
Storage**

**Data
Disposal**

DATA ANALYTICS **LIFECYCLE**



need to protect

**Data
Access**

**Data
Collection**

**Data
Preprocessing**

**Data
Sharing**

**Data
Storage**

**Data
Disposal**



DATA ANALYTICS **LIFECYCLE**



need to protect

**Data
Access**

**Data
Collection**

**Data
Preprocessing**

**Data
Sharing**

**Data
Storage**

**Data
Disposal**



DATA ANALYTICS **LIFECYCLE**



need to protect

**Data
Access**

**Data
Collection**

**Data
Preprocessing**

**Data
Sharing**

**Data
Storage**

**Data
Disposal**



DATA ANALYTICS **LIFECYCLE**



need to protect

**Data
Access**

**Data
Collection**

**Data
Preprocessing**

**Data
Sharing**

**Data
Storage**

**Data
Disposal**



DATA ACCESS

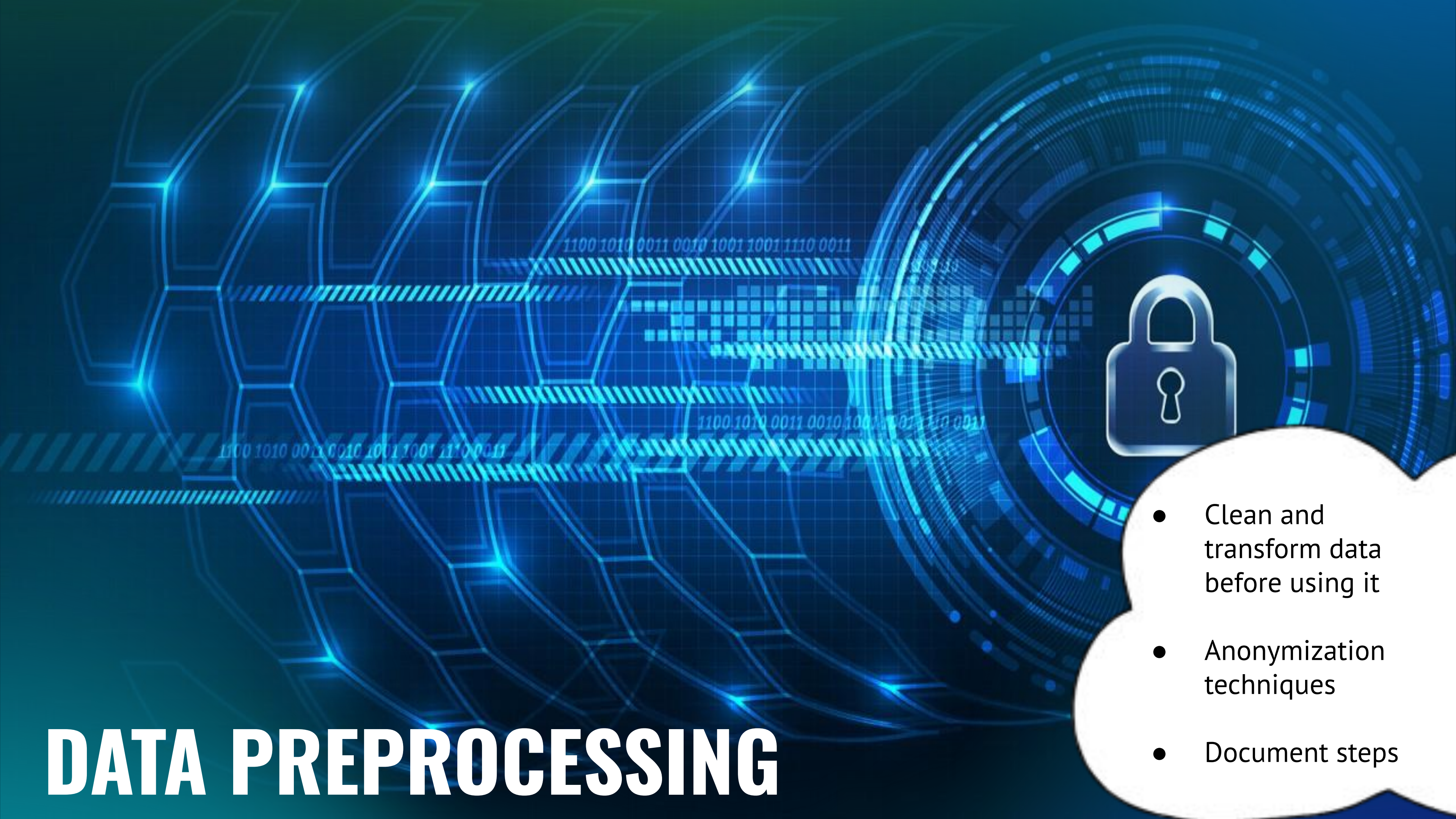


- Role Based Access (Principle of Least Privilege)
- Logical & Physical Access Control

data collection

- Minimize the amount of sensitive data you collect
- Avoid collecting unnecessary or irrelevant data

DATA PREPROCESSING

- 
- Clean and transform data before using it
 - Anonymization techniques
 - Document steps

DATA SHARING



- Only share with authorized and trusted parties
- Use secure and encrypted channels to transfer data

DATA STORAGE

- Encrypt data at rest and in transit
- Conduct regular backups

DATA DISPOSAL



- Delete the data in a timely manner when it is no longer needed
- Dispose of physical media or hardware properly

Why This Matters



Why This Matters



Data is EVERYWHERE!

We work with data every single day





***‘IT’S EASY TO STEAL
DATA THESE DAYS!’***





***‘IT’S EASY TO STEAL
DATA THESE DAYS!’***

~ Monica Kay Royal



CRIME-AS-A-SERVICE

According to Help Net Security, Crime-as-a-Service (CaaS) is the practice of experienced cybercriminals selling access to the tools and knowledge needed to execute cybercrime. CaaS enables cybercriminals to outsource various aspects of their operations, such as malware development, ransomware distribution, botnet rental, or data theft.

CRIME-AS-A-SERVICE

According to Help Net Security, Crime-as-a-Service (CaaS) is the practice of experienced cybercriminals selling access to the tools and knowledge needed to execute cybercrime. CaaS enables cybercriminals to outsource various aspects of their operations, such as malware development, ransomware distribution, botnet rental, or data theft.

The cost of global cybercrime has been estimated to reach

US \$10.5tn by 2025

Why This Matters



Data is EVERYWHERE!

Data is easy to steal.



Why This Matters



Data is EVERYWHERE!

Data is easy to steal.

Machine Learning models learn too much!!



Why This Matters



Data is EVERYWHERE!

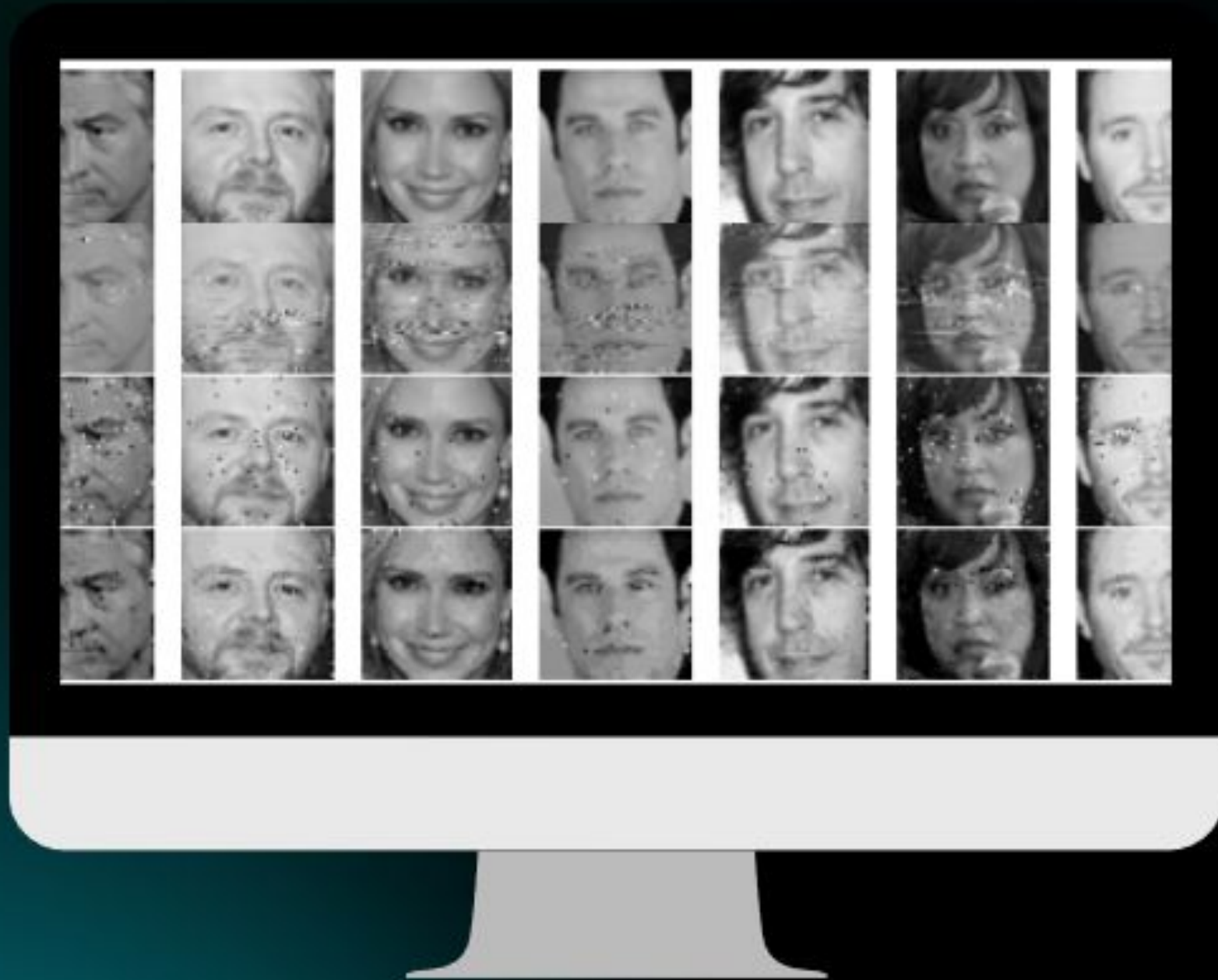
Data is easy to steal.

Machine Learning models learn too much!!

According to a paper from Cornell, ML models can learn too much from the training set which can be leaked and used to recreate part of the dataset if breached.



ML MODEL LEAKAGE



Trained the model using conventional training algorithms for image classification (row 1)

Experiments

- White-Box Attacks: adversary can inspect the parameters of the model after it is available
 - Correlated Value Encoding Attack (row 2)
 - Sign Encoding Attack (row 3)
- Black-Box Attacks: adversary can only send queries to a prediction API
 - Capacity Abuse Attack (row 4)



COUNTERMEASURES



Data Minimization

Only collect necessary data



Perturbation

Creating random noise



Synthetic Data

Create artificially generated data





DATA PREPROCESSING



ML MODELS



What is a model?



A program that can find patterns or make decisions from a previously unseen dataset.

Types of Decisions...

- Underwriting Process
- Targeted Advertising
- Hiring Process

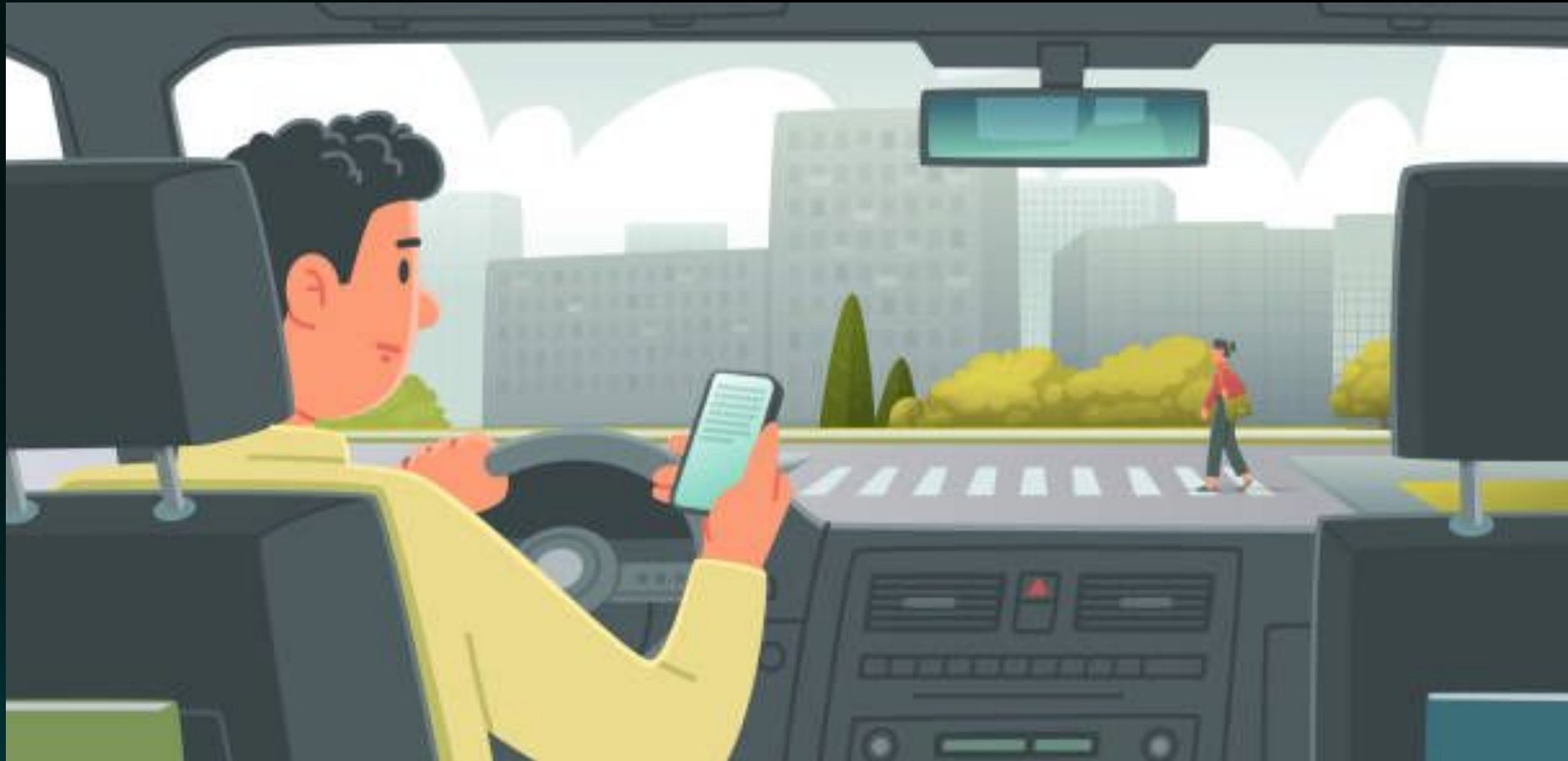




UNDERWRITING



UNDERWRITING



TARGETED ADS



TARGETED ADS



TARGETED ADS



EMPLOYMENT APPLICATION

Personal Information

Full Name (First, Middle, Last)

--	--	--

City/State/Zip Code

--	--	--

Phone Number

Email Address

--	--

Position & Availability

What position are you applying for?

Enter the date you can start work

--	--

TARGETED ADS



SPECIAL OFFER

Earn a \$200 Statement Credit and 20,000 Bonus Miles*

Plus enjoy your first checked bag free on Delta Flights

\$0 introductory annual fee for the first year, then \$99

*Statement credit issued approximately 8-12 weeks after you make a Delta purchase. Bonus miles will be issued after you make \$1,000 in purchases on your new Card in your first year.

[Rates & Fees](#)

[Offer Terms](#)



Apply for the Delta SkyMiles® Gold American Express Card

Southwest

[Already a Cardmember? >](#)

Select card below or [compare cards](#) >

Personal:



Southwest Rapid Rewards® Plus Credit Card

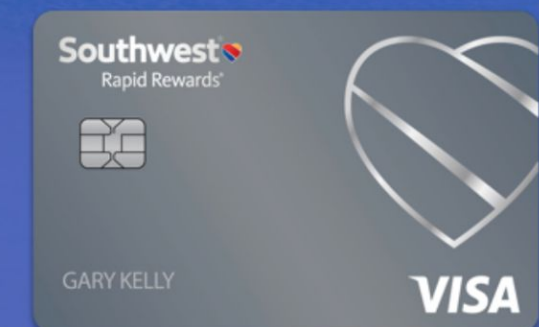
Earn 40,000 points

after you spend \$1,000 on purchases in the first 3 months from account opening.*

[Apply now >](#)

[Pricing & Terms](#) | [Offer Details](#)

\$69 annual fee applied to your first billing statement.†



(2,559 cardmember reviews)

HIRING PROCESS



HIRING PROCESS





Employment History

Purchasing Patterns

Personality Tests

Personal Income

Demographics

Credit Scores

Locations

Behaviors

GPA

ANONYMIZATION TECHNIQUES



Mask

Concealing specific parts of data



Generalization

Aggregating data at a higher level



Pseudonymization

Creating pseudo names to protect identities

Randomization

Slightly modifying dates

Data Swapping

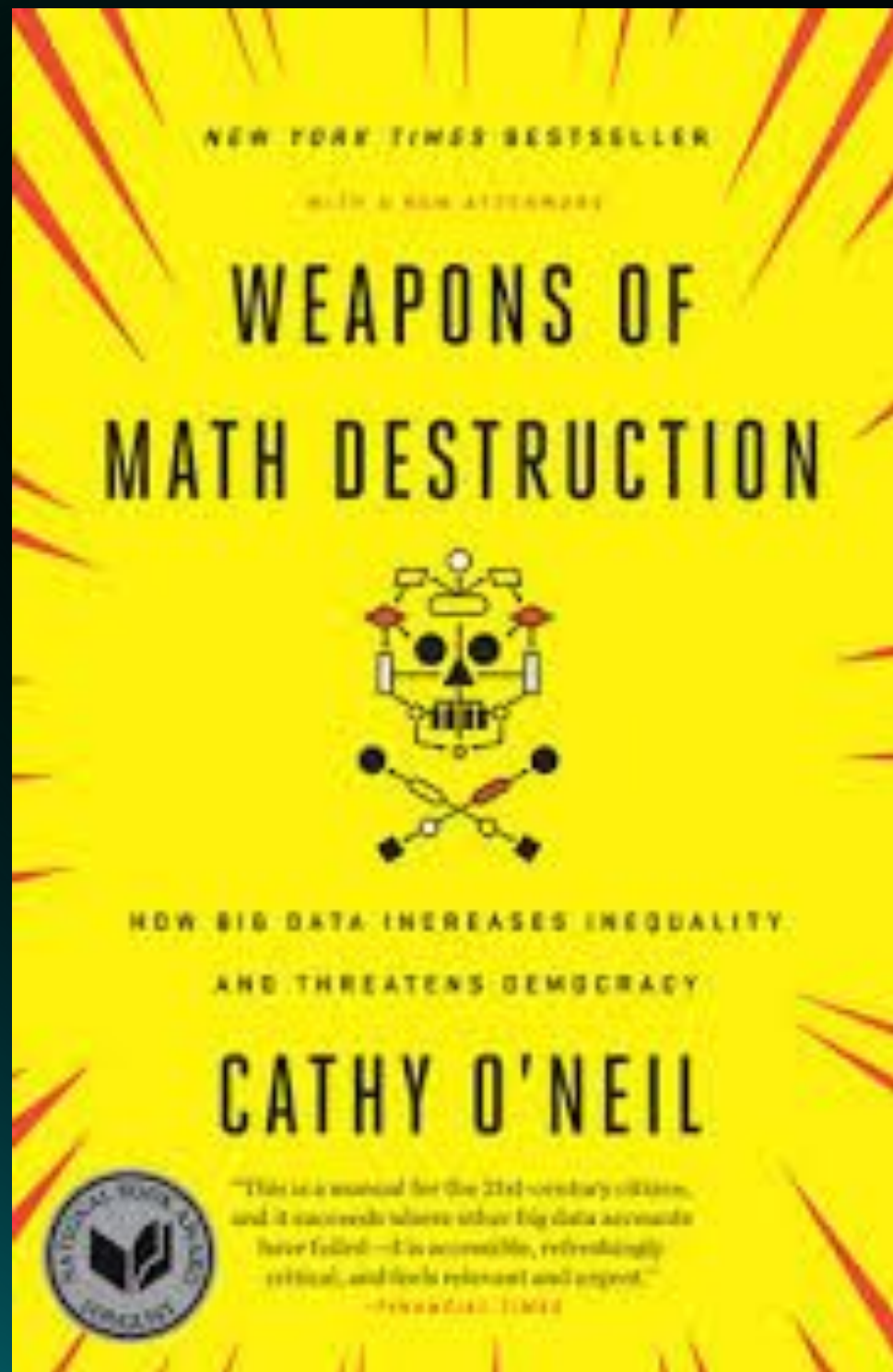
Shuffling specific attributes

Tokenization

Replacing sensitive data with unique identifiers



DISARM WMDs



DISARM WMDs

- Reevaluate Success

SUCCESS =



DISARM WMDs

- Reevaluate Success
- Feedback Loops



DISARM WMDs

- Reevaluate Success
- Feedback Loops
- Transparency



DISARM WMDs

- Reevaluate Success
- Feedback Loops
- Transparency
- Regulation Updates

Fair Credit Reporting Act (FCRA)

Equal Credit Opportunity Act (ECOA)

Americans with Disabilities Act (ADA)

DISARM WMDs

- Reevaluate Success
- Feedback Loops
- Transparency
- Regulation Updates
- Algorithmic Audits



FUTURE TRENDS

- Privacy-Preserving AI / ML Frameworks
- Advanced Anonymization Techniques
- Cross-disciplinary Approaches
- Educational Initiatives



SYNTHETIC DATA





Faker

[Faker](#) is a Python package that generates fake data for you. Installation: [Help Link](#) Open Anaconda prompt command to install:

```
conda install -c conda-forge faker
```

Import package

```
from faker import Faker
```

Faker has the ability to print/get a lot of different fake data, for instance, it can print fake name, address, email, text, etc.

Important most commonly used faker commands

```
fake.name()  
fake.address()  
fake.email()  
fake.text()  
fake.country()
```




Faker

OUTPUT:(Different every time)

vwilson@hotmail.com

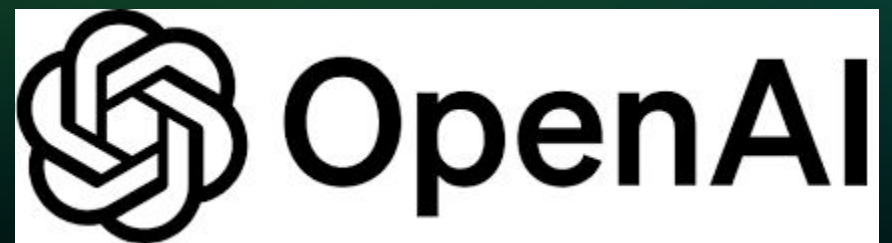
Belgium

Shane Hunter

Commodi vel libero placeat quibusdam odio odio consequatur. Ducimus libero quae optio non quidem. Facilis
quas impedit quo.

26.5687745 -124.802165

<http://www.turner.com/>



Getting setup

```
%pip install openai  
%pip install pandas  
%pip install scikit-learn  
%pip install matplotlib
```



```
from openai import OpenAI  
import os  
import re  
import numpy as np  
import pandas as pd  
from sklearn.cluster import KMeans  
import matplotlib.pyplot as plt  
import json  
import matplotlib
```



```
client = OpenAI(api_key=os.environ.get("OPENAI_API_KEY", "<your OpenAI API key if not set as env var>"))
```




1. CSV with a structure prompt [↗](#)

Here we create data in the simplest way. You can quickly generate data by addressing 3 key points: telling it the format of the data (CSV), the schema, and useful information regarding how columns relate (the LLM will be able to deduce this from the column names but a helping hand will improve performance).



```
datagen_model = "gpt-4o-mini"
question = """
Create a CSV file with 10 rows of housing data.
Each row should include the following fields:
- id (incrementing integer starting at 1)
- house size (m^2)
- house price
- location
- number of bedrooms
```

Make sure that the numbers make sense (i.e. more rooms is usually bigger size, more expensive locations increase price. more size is usually higher price etc. make sure all the numbers make sense). Also only respond with the CSV.

```
"""

response = client.chat.completions.create(
    model=datagen_model,
    messages=[
        {"role": "system", "content": "You are a helpful assistant designed to generate synthetic data."},
        {"role": "user", "content": question}
    ]
)
res = response.choices[0].message.content
print(res)
```




```
```csv
id,house_size_m2,house_price,location,number_of_bedrooms
1,50,150000,Suburban,2
2,75,250000,City Center,3
3,100,350000,Suburban,4
4,120,450000,Suburban,4
5,80,300000,City Center,3
6,90,400000,City Center,3
7,150,600000,Premium Area,5
8,200,750000,Premium Area,5
9,55,180000,Suburban,2
10,300,950000,Premium Area,6
```
```




```
datagen_model = "gpt-4o-mini"
question = """
Create a CSV file with 10 rows of housing data.
Each row should include the following fields:
- id (incrementing integer starting at 1)
- house size (m^2)
- house price
- location
- number of bedrooms
```

Make sure that the numbers make sense (i.e. more rooms is usually bigger size, more expensive locations increase price. more size is usually higher price etc. make sure all the numbers make sense). Also only respond with the CSV.

```
"""

response = client.chat.completions.create(
    model=datagen_model,
    messages=[
        {"role": "system", "content": "You are a helpful assistant designed to generate synthetic data."},
        {"role": "user", "content": question}
    ]
)
res = response.choices[0].message.content
print(res)
```




```
output_string = ""
for i in range(3):
    question = f"""
I am creating input output training pairs to fine tune my gpt model. The usecase is a retailer generating a description
The format should be of the form:
1.
Input: product_name, category
Output: description
2.
Input: product_name, category
Output: description

Do not add any extra characters around that formatting as it will make the output parsing break.
Create as many training pairs as possible.
"""

    response = client.chat.completions.create(
        model=datagen_model,
        messages=[
            {"role": "system", "content": "You are a helpful assistant designed to generate synthetic data."},
            {"role": "user", "content": question}
        ]
    )
    res = response.choices[0].message.content
    output_string += res + "\n" + "\n"
print(output_string[:1000]) #displaying truncated response
```

I am creating input output training pairs to fine tune my gpt model. The usecase is a retailer generating a description for a product from a product catalogue. I want the input to be product name and category (to which the product belongs to) and output to be description.



1.

Input: Wireless Bluetooth Headphones, Electronics

Output: Immerse yourself in high-quality sound with these Wireless Bluetooth Headphones, featuring active noise cancellation and a comfortable over-ear design for extended listening sessions.

2.

Input: Organic Green Tea, Beverages

Output: Enjoy a refreshing cup of Organic Green Tea, sourced from the finest leaves, packed with antioxidants, and perfect for a healthy, invigorating boost anytime.

3.

Input: Stainless Steel Kitchen Knife, Kitchenware

Output: Cut with precision and ease using this Stainless Steel Kitchen Knife, designed with an ergonomic handle and a sharp blade for all your culinary tasks.

4.

Input: Hiking Backpack, Outdoor Gear

Output: Explore the great outdoors with this durable Hiking Backpack, featuring multiple compartments for optimal organization and a breathable design for ultimate comfort on long treks.

5.

Input: Air Fryer, Kitchen Appliances

Output: Cook your favorite meals with less oil using this Air Fryer

SYNTHETIC DATA



PANEL:

USE CASES AND SECURITY CONSIDERATIONS FOR GENAI



Marie Jadhvani



Pete Stiglich

Protecting Data for Analysis and Machine Learning



RESOURCES

- O'Neil, C. (2016). Weapons of Math Destruction, How Big Data Increases Inequality and Threatens Democracy. Penguin Books
- Song, C. Ristenpart, T. Shmatikov, V. (2017). Machine Learning Models that Remember Too Much. Association for Computing Machinery.
- Python Faker Library: <https://www.geeksforgeeks.org/python/python-faker-library/>
- OpenAI Cookbook: Senthetic data generation (Part 1):
<https://cookbook.openai.com/examples/sdg1>



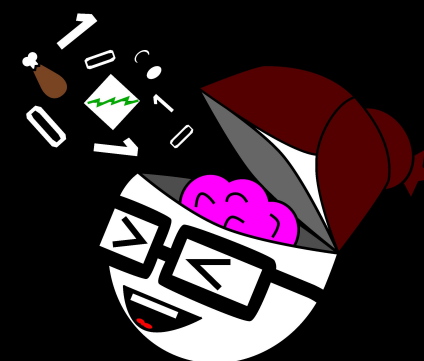
<https://www.linkedin.com/in/monicakayroyal/>



www.nerdnourishment.com



[@nerdnourishment](https://www.youtube.com/@nerdnourishment)



Protecting Data for Analysis
and Machine Learning

