

1

Official (Open)

## Welcome and Introduction

- **Outline**
  - **Introduction to Generative AI**
  - **Primer on Data Science**
  - **Generative AI Assistance for Data Science**

Gen AI Assisted Data Science

2

2

1

Official (Open)

## Notes and Updates

- <https://peterleong.github.io/gads/>

Gen AI Assisted Data Science

3

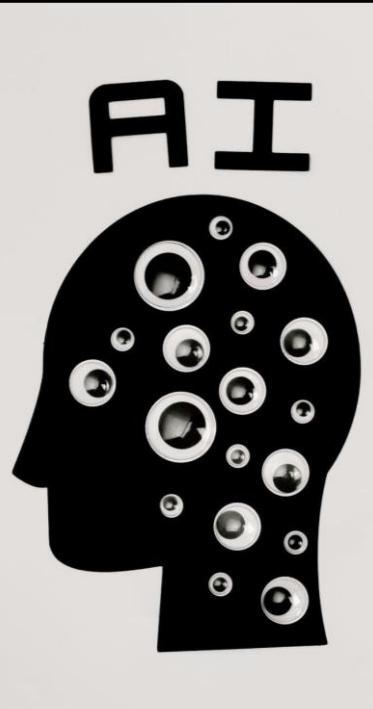
3

Official (Closed), Sensitive (Normal)

## SKILLSFUTURE CLAIMABLE COURSES

4

2



## SKILLSFUTURE SHORT COURSES

-  [Introduction to AI and Machine Learning](#)
-  [Data Visualization with Tableau](#)
-  [Generative AI and ChatGPT for Business](#)

5

## SKILLSFUTURE SPECIALIST DIPLOMAS

-  [Specialist Diploma in Data Science \(Artificial Intelligence\)](#)
-  [Diploma \(Conversion\) in Web & Programming](#)
-  [Specialist Diploma in Data Protection and AI Governance](#)
-  [Diploma in Infocomm & Digital Media \(Cyber Security\)](#)
-  [Specialist Diploma in Cloud Computing \(DevOps\)](#)
-  [Specialist Diploma in Operational Technology Cybersecurity \(Ethical Hacking and Incident Response\)](#)
-  [Specialist Diploma in Full Stack Web Development](#)

6



7

Official (Open)

# Generative AI isn't about what you think it is

Gen AI Assisted Data Science

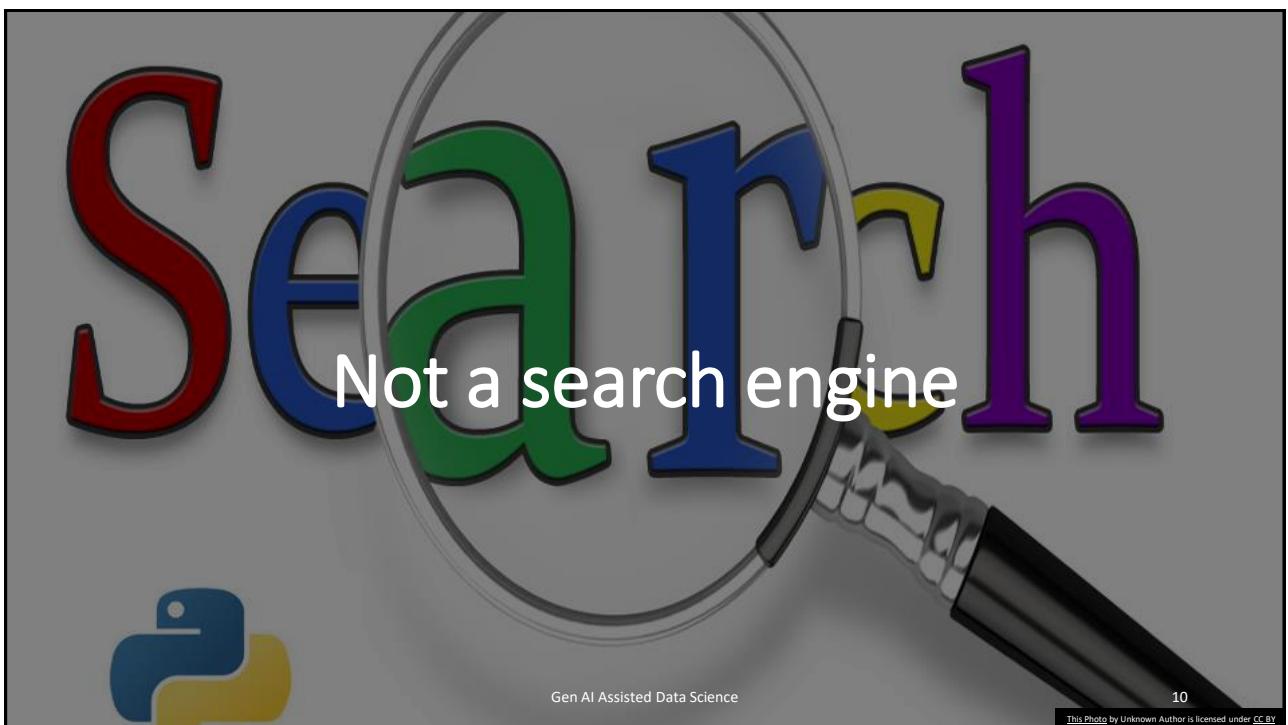
8

8

4



9



10



**Not a new  
way of  
programming**

11



12

Official (Open)

# What then

Let's get everyone's input  
in a word cloud

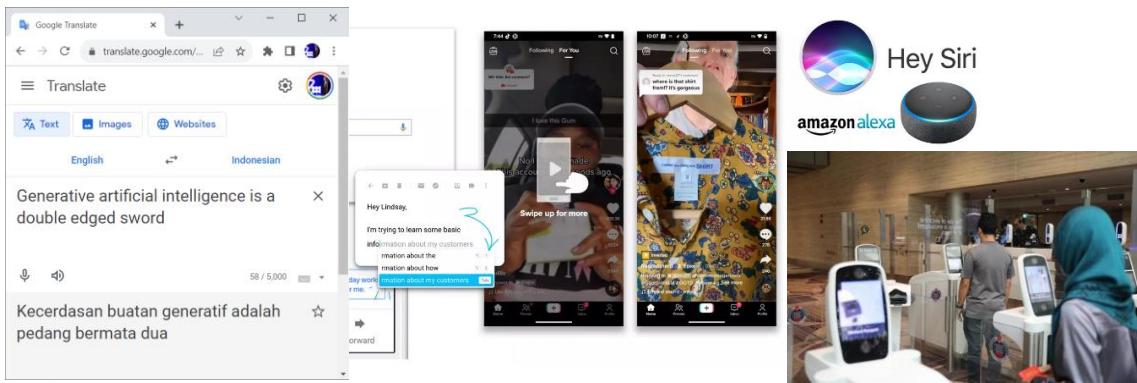
Gen AI Assisted Data Science

13

13

Official (Open)

AI has been with us for years, whether  
“generative” or not



Can you think of more AI examples in our daily lives?

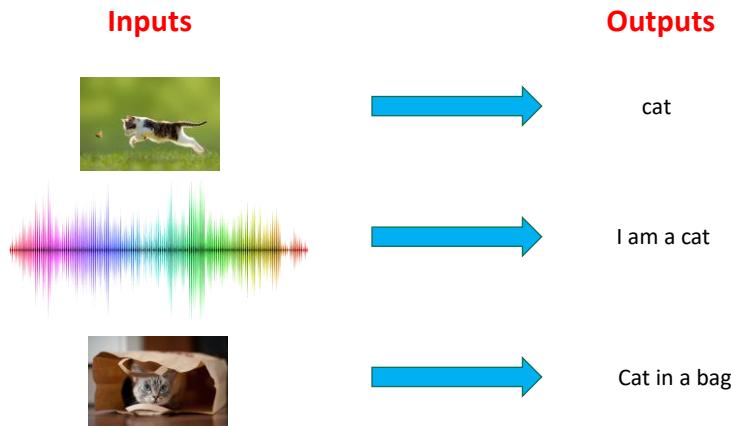
Gen AI Assisted Data Science

14

14

Official (Open)

## AI BEFORE



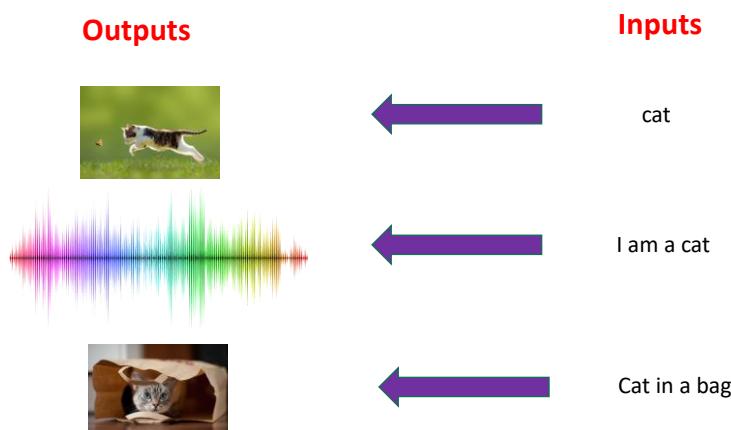
Gen AI Assisted Data Science

15

15

Official (Open)

## Generative AI – The AI Inverse



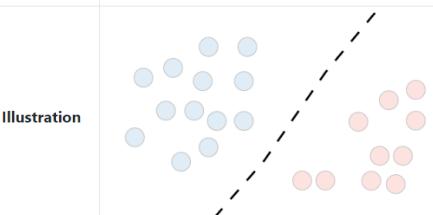
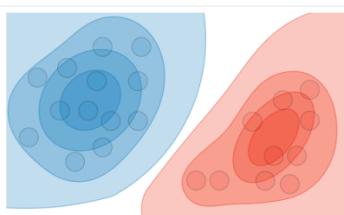
Gen AI Assisted Data Science

16

16

Official (Open)

## Discriminative model vs Generative model

	Discriminative model	Generative model
Goal	Directly estimate $P(y x)$	Estimate $P(x y)$ to then deduce $P(y x)$
What's learned	Decision boundary	Probability distributions of the data
Illustration		
Examples	Regressions, SVMs	GDA, Naive Bayes

Source: <https://stanford.edu/~shervine/teaching/cs-229/cheatsheet-supervised-learning>

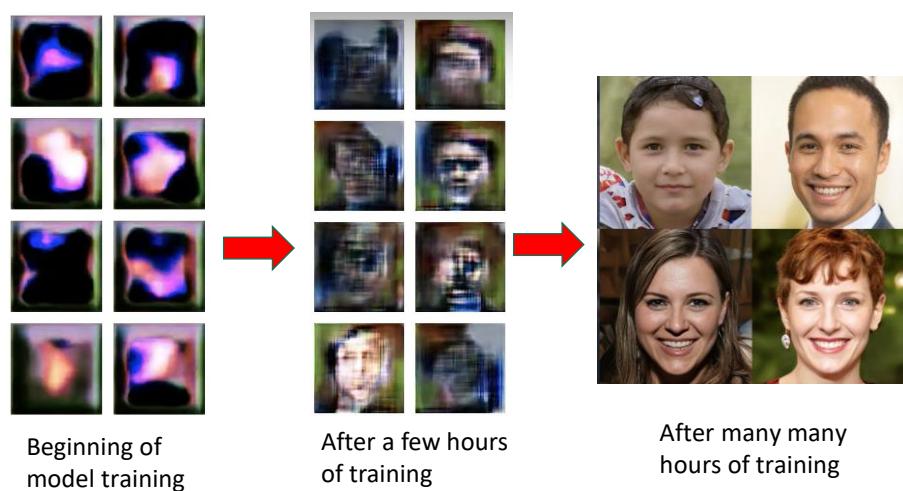
Gen AI Assisted Data Science

17

17

Official (Open)

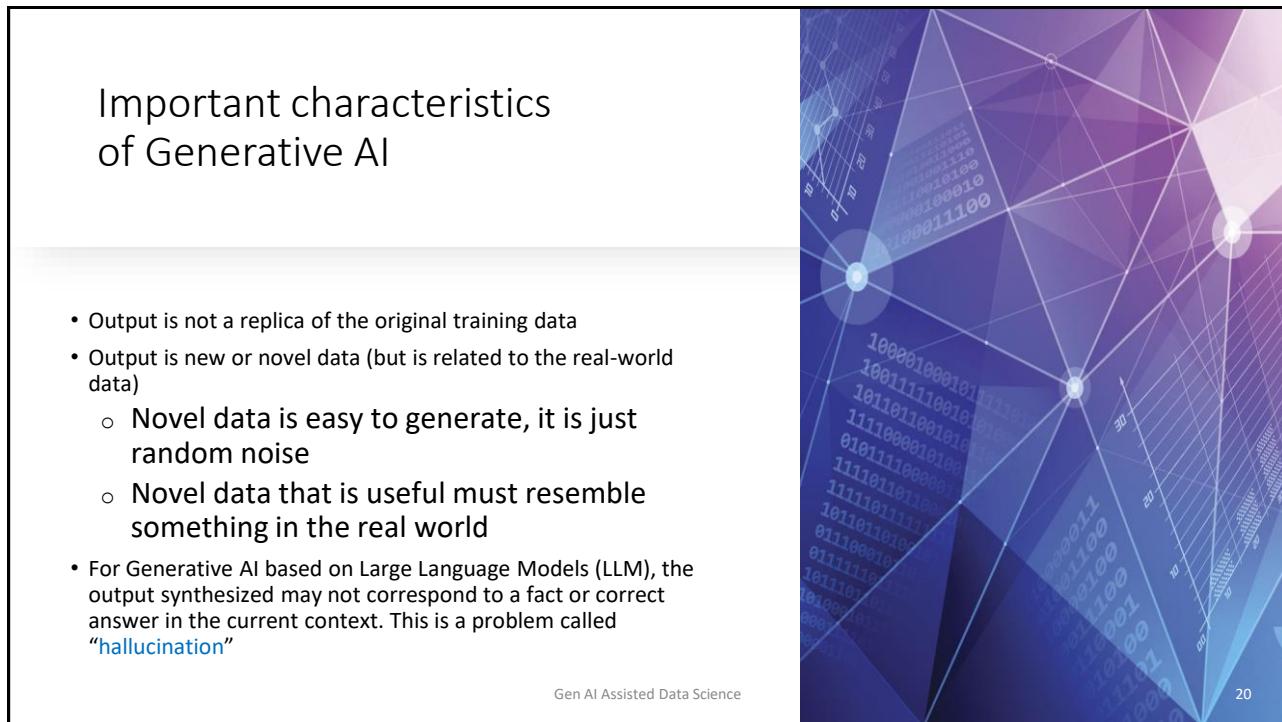
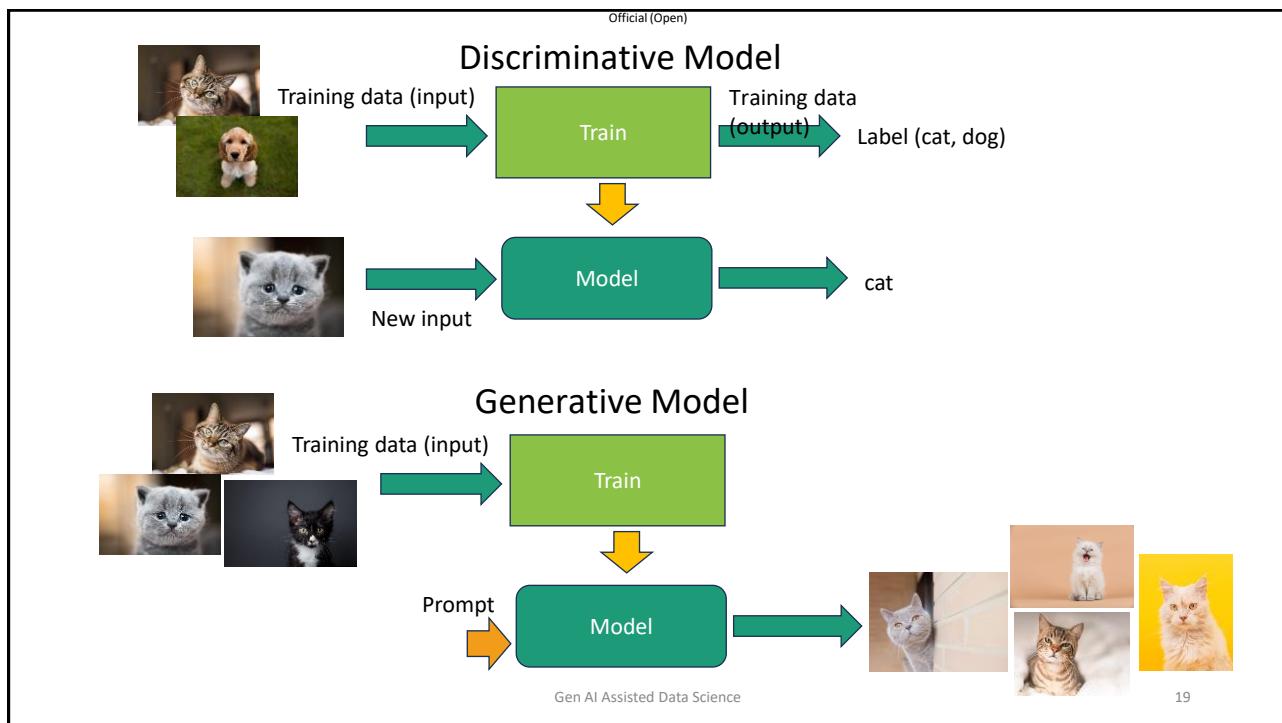
## How AI learn to generate human face



Gen AI Assisted Data Science

18

18



Official (Open)

# What are some use cases?

Let's get everyone's input  
in a word cloud

Gen AI Assisted Data Science

21

21

## Impact of Generative AI

Output Type	Potential Impact	Examples
Text	- Generate realistic and creative text formats, including poems, code, scripts, musical pieces, email, letters, etc.	- Write news articles, poems, and scripts - Translate languages - Generate marketing copy - Create email and letters - Summarize documents
Code	- Generate code for various programming languages - Debug and optimize existing code - Automate repetitive coding tasks - Develop new algorithms and machine learning models	- Generate code for websites, mobile apps, and games - Debug and optimize code for performance - Automate code testing and deployment - Develop new AI algorithms and machine learning models
Images	- Create realistic and high-quality images from text descriptions - Generate creative and artistic images - Enhance and restore existing images - Design graphics and illustrations	- Generate images for product mockups, social media posts, and presentations - Create artistic images and illustrations - Enhance and restore old photos - Design graphics for websites and marketing materials
Audio	- Generate music and sound effects - Compose songs and melodies - Create soundtracks for movies and video games - Convert text to speech - Synthesize realistic voices	- Generate music for background music and sound effects - Compose songs and melodies for singers and musicians - Create soundtracks for movies, video games, and advertisements - Convert text to speech for audiobooks and e-learning materials - Synthesize realistic voices for virtual assistants and chatbots
Video	- Generate realistic and high-quality videos from text descriptions - Create animated videos and special effects - Edit and manipulate existing videos - Create short films and documentaries	- Generate videos for product demos, explainer videos, and social media posts - Create animated videos and special effects for movies and video games - Edit and manipulate existing videos for marketing and entertainment purposes - Create short films and documentaries for educational and entertainment purposes
3D Objects	- Generate 3D models from text descriptions - Design and create 3D models for products, buildings, and characters - Print 3D models for prototyping and manufacturing	- Generate 3D models for product design and prototyping - Design and create 3D models for architecture and engineering - Create 3D models for characters and animation
Biology	- Generate new DNA sequences - Design new drugs and therapies - Predict protein structures and functions - Develop new diagnostic and treatment methods	- Develop new vaccines and antibiotics - Design new drugs for cancer and other diseases - Predict protein structures and functions for drug discovery - Develop new diagnostic methods for diseases
Materials	- Design new materials with desired properties - Discover new materials for various applications - Develop new manufacturing processes for materials	- Design new materials for energy storage, batteries, and solar cells - Discover new materials for aerospace, automotive, and medical applications - Develop new manufacturing processes for materials that are more efficient and environmentally friendly

22

## Implications & Risks

### Implications

- The Generative AI model may learn information that is copyrighted (e.g. logos)
- The Generative AI model may copy the style of living artists and writers (copycat artist)
- Generative AI can widen the digital divide
- Jobs disruptions for tasks that may be automated

### Risks

- No explicit content given for your private data (e.g. your photo) to be used for training data
- Confidential data being leaked when using publicly accessible Generative AI
- Naïve users may not realize that model is hallucinating because the answer is very convincing
- Generative AI can be used to create disinformation or adversarial examples to fool AI
- Biased outputs from Generative AI

Gen AI Assisted Data Science

23

23

## Examples of bias in Generative Models

(Demo)

Gen AI Assisted Data Science

24

24

12

Official (Open)

Create a photo realistic picture of delicious steak dinner



Gen AI Assisted Data Science

25

25

Official (Open)

Create a photo realistic image of delicious hainanese chicken rice



Gen AI Assisted Data Science

26

26

13

Official (Open)

please create a photo realistic picture of a bowl of delicious Singapore fish head curry



Gen AI Assisted Data Science

27

27

## Section 2: Primer on Data Science

Gen AI Assisted Data Science

28

28

14

Official (Open)

**Data science is the art of extracting knowledge from data, using a blend of statistics, computer science, and mathematics to uncover hidden patterns, predict the future, and make better decisions.**

**It's like sifting through sand to find hidden gems, turning information into power for businesses and organizations.**



Gen AI Assisted Data Science

29

29

## Overview of Data Science Process / Workflow

Gen AI Assisted Data Science

30

30

Official (Open)

## Data Collection

- Sources of data (structured, unstructured, APIs, databases)
- Data quality assessment and cleaning
- Sampling techniques

Gen AI Assisted Data Science

31

31

Official (Open)

## Data Exploration and Preprocessing:

- Descriptive statistics (mean, median, variance)
- Visualization (histograms, scatter plots, heatmaps)
- Handling missing values and outliers

Gen AI Assisted Data Science

32

32

Official (Open)

## Feature Engineering

- **Creating relevant features from raw data**
- **Techniques (binning, scaling, one-hot encoding)**
- **Domain-specific feature extraction**

Gen AI Assisted Data Science

33

33

Official (Open)

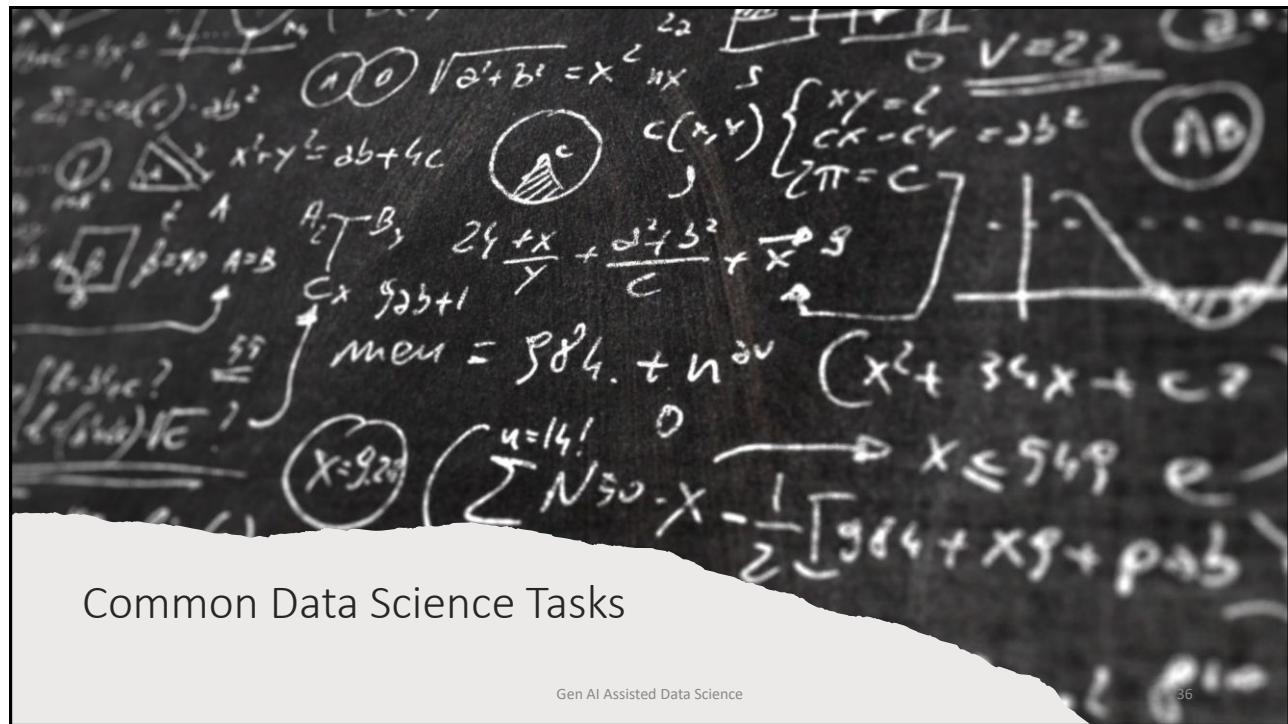
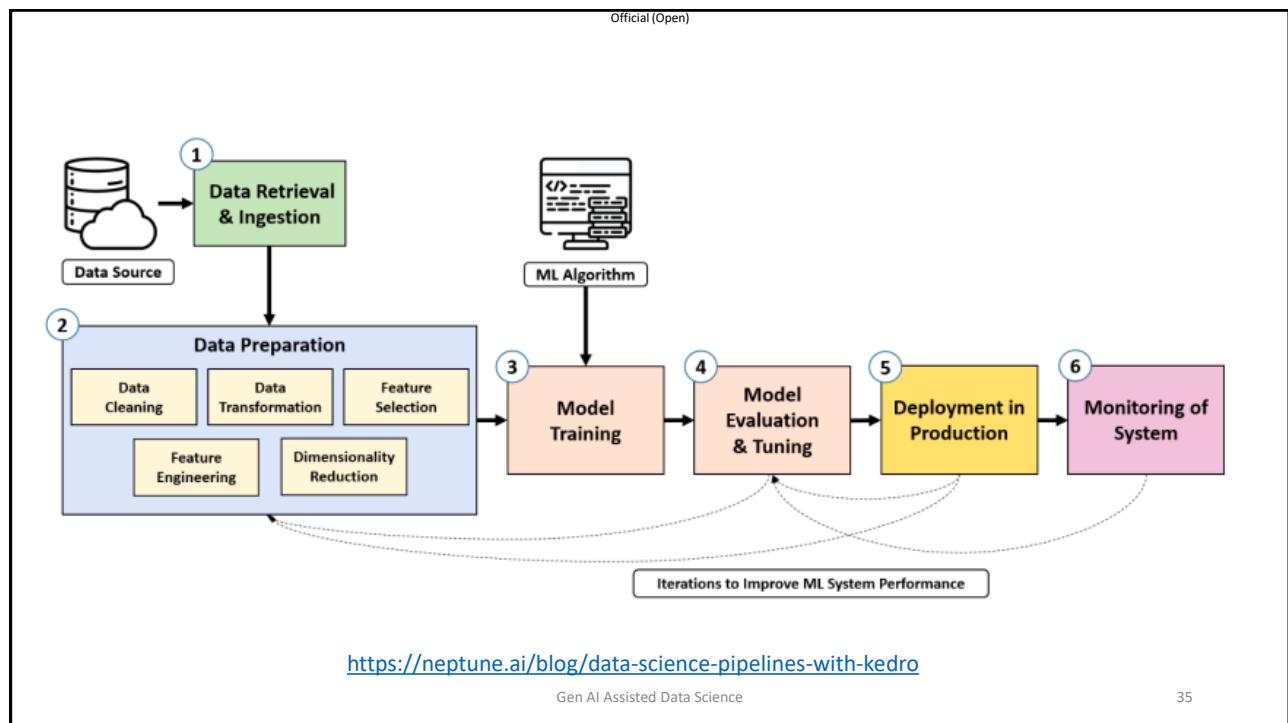
## Model Selection and Evaluation

- **Choosing appropriate algorithms (linear regression, decision trees, neural networks).**
- **Cross-validation and hyperparameter tuning.**
- **Evaluation metrics (accuracy, precision, recall).**

Gen AI Assisted Data Science

34

34



Official (Open)

# Regression

- Predicting continuous numeric values (e.g., house prices, stock prices)
  - Linear regression, polynomial regression



37

37

Official (Open)

# Classification

- Categorizing data into classes (e.g., spam vs. non-spam emails, disease diagnosis)
  - Logistic regression, decision trees, support vector machines



38

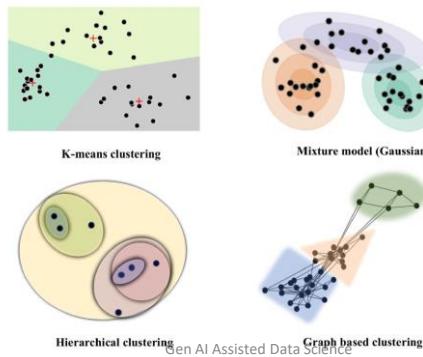
38

19

Official (Open)

# Clustering

- Grouping similar data points (e.g., customer segmentation, image segmentation)
  - K-means, hierarchical clustering



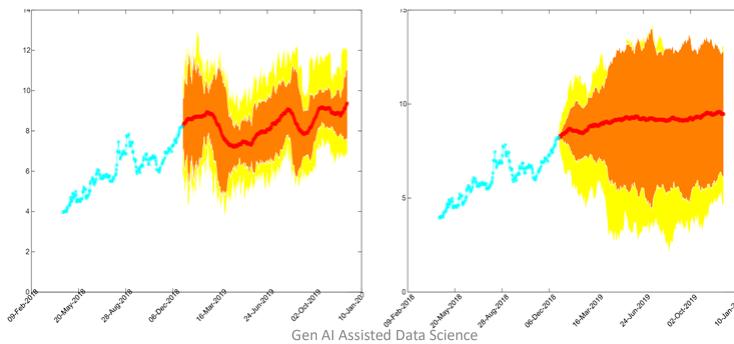
39

39

Official (Open)

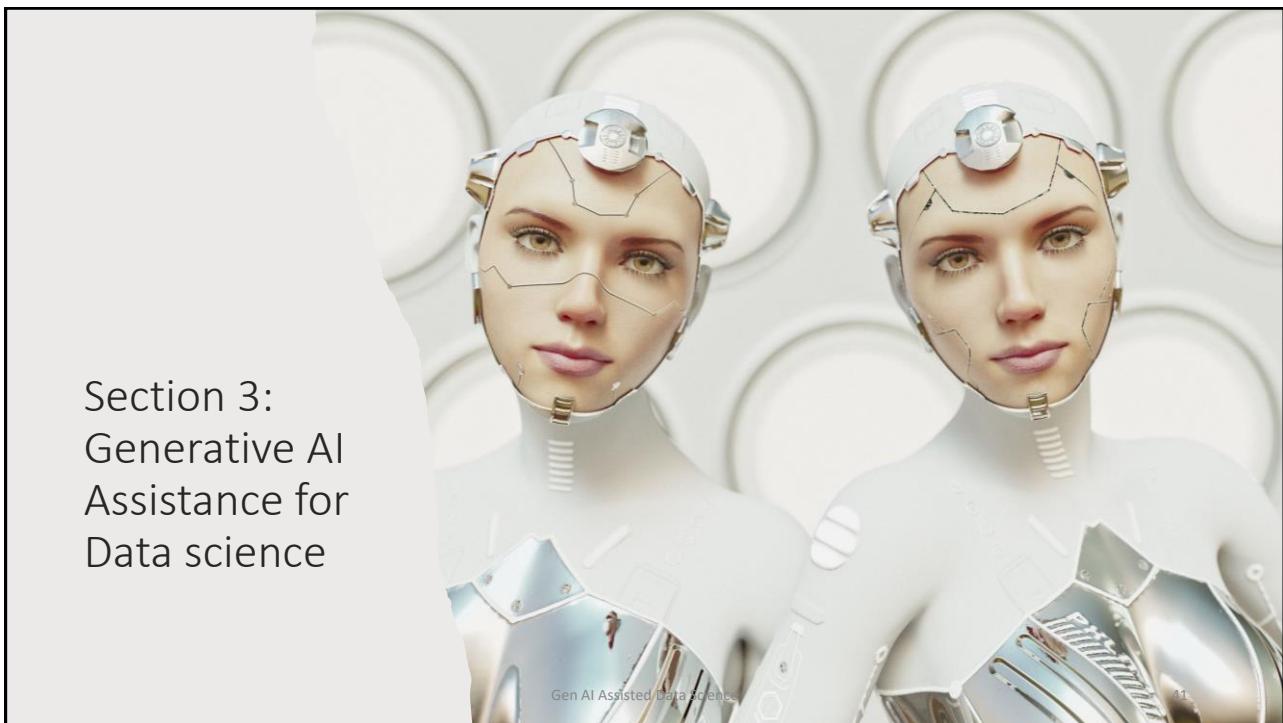
# Time Series Analysis

- Analyzing data with temporal dependencies (e.g., stock prices, weather data)
  - Autoregressive models, moving averages



40

40



### Section 3: Generative AI Assistance for Data science

41



## Generative Models in Data Science

Gen AI Assisted Data Science

42

42



# Interactive Activity

Gen AI Assisted Data Science

43

Official (Open)

## Let's ask ChatGPT to create some data

Prompt: create a synthetic data set for visits to [www.mywebsite.com](http://www.mywebsite.com)

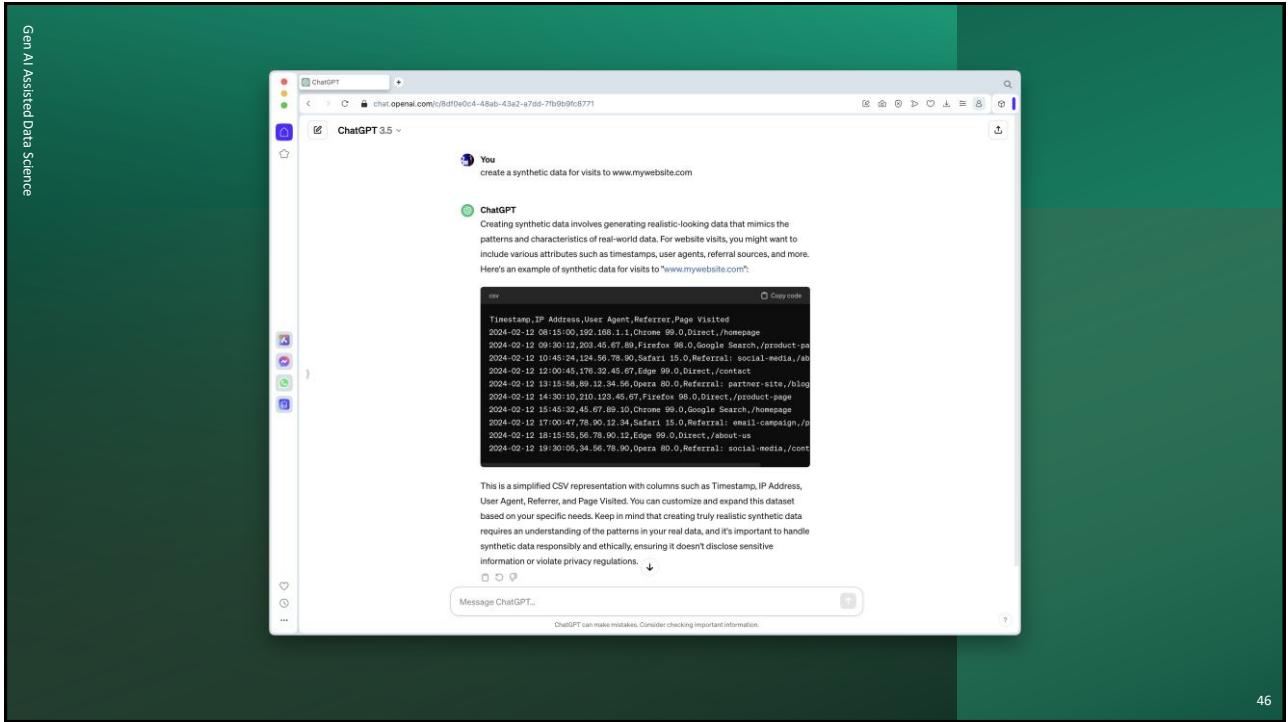
Gen AI Assisted Data Science

44



45

45



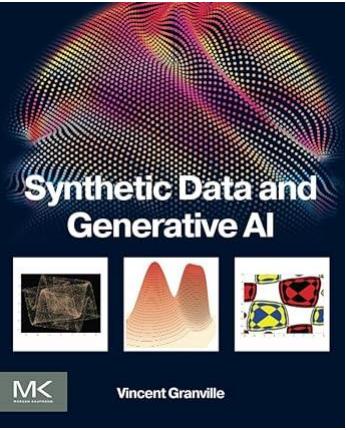
46

46



47

Official (Open)



*Synthetic Data and Generative AI* covers the foundations of machine learning with modern approaches to solving complex problems and the systematic generation and use of synthetic data. Emphasis is on scalability, automation, testing, optimizing, and interpretability (explainable AI). For instance, regression techniques – including logistic and Lasso – are presented as a single method without using advanced linear algebra. Confidence regions and prediction intervals are built using parametric bootstrap without statistical models or probability distributions. Models (including generative models and mixtures) are mostly used to create rich synthetic data to test and benchmark various methods. Emphasizes numerical stability and performance of algorithms (computational complexity)

- Focuses on explainable AI/interpretable machine learning, with heavy use of synthetic data and generative models, a new trend in the field
- Includes new, easier construction of confidence regions, without statistics, a simple alternative to the powerful, well-known XGBoost technique
- Covers automation of data cleaning, favoring easier solutions when possible
- Includes chapters dedicated fully to synthetic data applications: fractal-like terrain generation with the diamond-square algorithm, and synthetic star clusters evolving over time and bound by gravity

Source: <https://www.amazon.com/Synthetic-Data-Generative-Vincent-Granville/dp/0443218579>

Gen AI Assisted Data Science

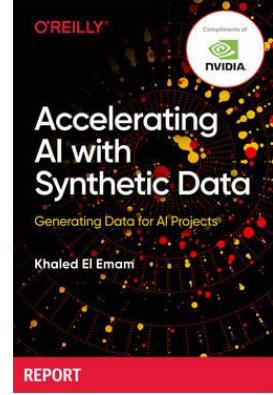
48

48

Official (Open)

# Synthetic Data Generation

- Using generative models to create realistic synthetic data.
  - Addressing data scarcity issues.

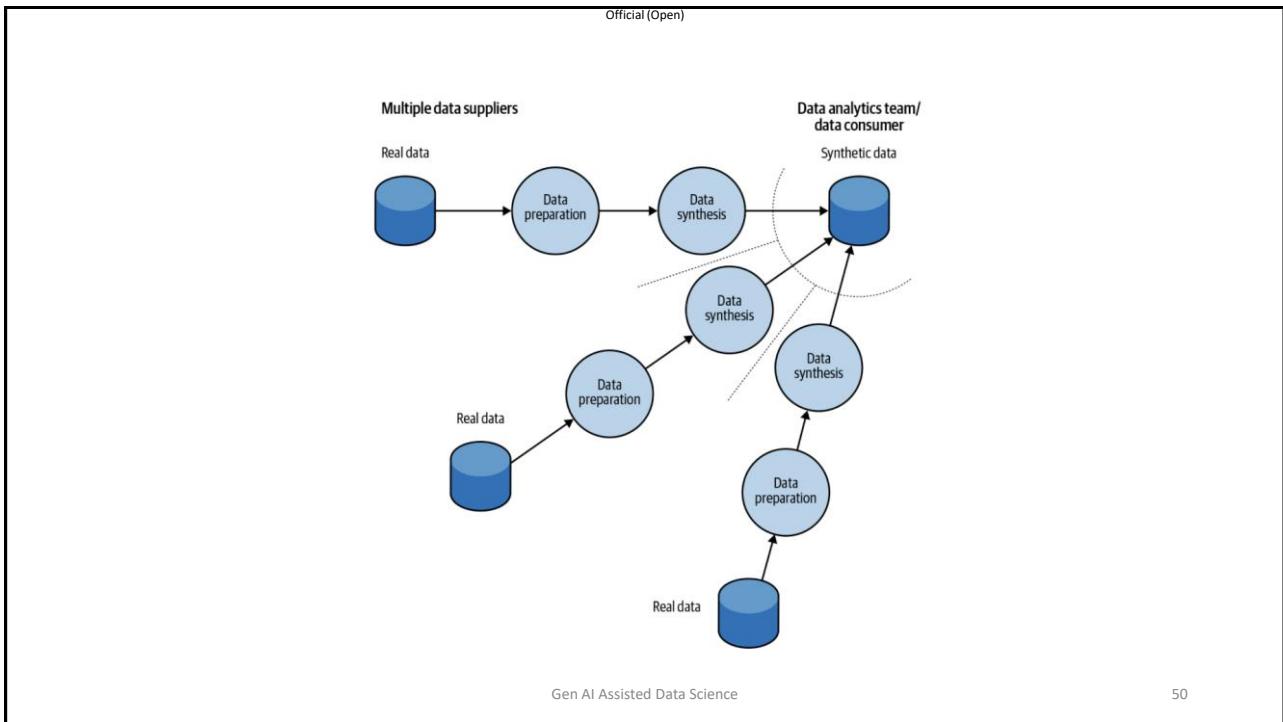


**O'REILLY®**  
Compliments of  
**NVIDIA**  
**Accelerating AI with Synthetic Data**  
Generating Data for AI Projects®  
Khaled El Emam  
**REPORT**

Source: <https://www.oreilly.com/library/view/accelerating-ai-with/9781492045991/>

Gen AI Assisted Data Science

49



Official (Open)

## Anomaly Detection

- **Identifying unusual patterns in data (fraud detection, network intrusion).**
  - Autoencoders, variational autoencoders.

Gen AI Assisted Data Science

51

51

Official (Open)

Generative AI models can create synthetic data that mimics legitimate transactions. By training on both genuine and synthetic data, models become adept at detecting unusual patterns indicative of fraud.

Adaptability: Unlike rule-based systems, Generative AI models can adapt to new fraud patterns.

Gen AI Assisted Data Science

52

52

Official (Open)

# How does Generative AI detect and fix Data Anomalies?

## Data Anomaly Detection

- **Autoencoders:** Utilize autoencoder neural networks, an unsupervised learning algorithm, to learn the representation of routine data. Anomalies can then be detected by identifying data points that do not reconstruct well.
- **GANs (Generative Adversarial Networks):** Train GANs to generate synthetic data like the actual data. Any deviation between accurate data and generated data can indicate anomalies.
- **Variational Autoencoders (VAEs):** VAEs can learn the underlying distribution of regular data and identify data points that do not conform to this distribution as anomalies.

Source: <https://www.xenonstack.com/blog/use-of-generative-ai-to-detect-and-fix-data-anomalies>

Gen AI Assisted Data Science

53

53

Official (Open)

# How does Generative AI detect and fix Data Anomalies?

## Data Anomaly Fixing

- **Imputation Techniques:** Use generative models to impute missing or anomalous data points. Generative models can generate plausible values for missing data points based on the learned patterns in the dataset.
- **Data Synthesis:** Generate synthetic data to replace anomalous data points. This is especially useful when dealing with sensitive data, as you can maintain privacy while ensuring the dataset's quality.
- **Conditional GANs:** Train conditional GANs where the condition is the surrounding data context. Generate data points conditioned on their neighboring points, ensuring coherence and realism.

Source: <https://www.xenonstack.com/blog/use-of-generative-ai-to-detect-and-fix-data-anomalies>

Gen AI Assisted Data Science

54

54

Official (Open)

## Activity:

- <https://taylordenouden.com/VAE-Latent-Space-Explorer/>

VAE Latent Space Explorer

Mu: 0.012  
Sigma: 0.328

Gen AI Assisted Data Science

55

55

Official (Open)

# Gretel.ai

Welcome to Gretel!

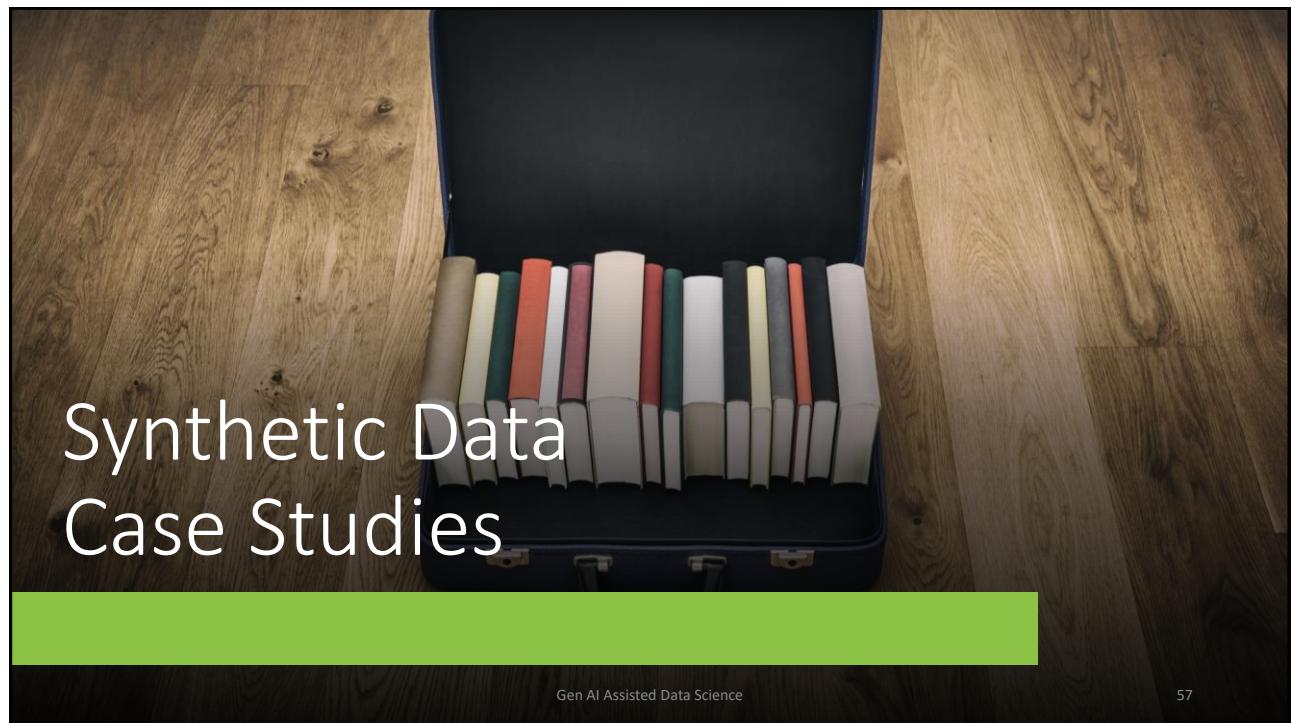
The developer platform for synthetic data.

With Gretel, developers can get started in minutes with open source reference examples and simple APIs for generating unlimited amounts of **synthetic data**, labeling personally identifiable information, or anonymizing and removing biases from data. Gretel services are controlled by a **simple web-based interface** and run in Gretel's managed cloud service or within your own private cloud environment.

Gen AI Assisted Data Science

56

56



# Synthetic Data Case Studies

Gen AI Assisted Data Science

57

57

Official (Open)

## Data Augmentation

- Enhancing model performance by augmenting the training dataset
- Applying generative models for data expansion

Gen AI Assisted Data Science

58

58

29

# Why Generative AI for Data Science?

Gen AI Assisted Data Science

59

59

Official (Open)

## Efficiency Gains

- **Generative AI streamlines repetitive tasks (e.g., data augmentation, content creation).**
- **Time savings for data scientists.**

For example, an analyst who is a researcher could use their exploratory models on synthetic data to then apply for funding to get access to the real data, which may require a full protocol and multiple levels of approvals. In such an instance, work with synthetic data that does not produce good models or actionable results would still be beneficial because analysts would have avoided the extra effort required to get access to the real data for a potentially futile analysis.

Source: <https://www.oreilly.com/library/view/accelerating-ai-with/9781492045991/>

Gen AI Assisted Data Science

60

60

30

Official (Open)

# Productivity Boost

- Examples of generative models accelerating model development.
- Impact on project timelines.

The image shows two side-by-side screenshots of the ChatGPT 3.5 web interface. Both windows have the title 'ChatGPT 3.5' and show the URL 'https://chat.openai.com/c/fc2902a0-cf94-4ec4-9a1e-e1ddee0f390a0'. The left window shows a user message 'You create code for a linear mixed effects model using python. the y is the predicted score, the X is the subject, medianscore, attendance, worstscorescore, use subject as random effect and worstscorescore as random intercept. subject is a categorical variable, others are numeric.' Below it is a code block:

```

python
import pandas as pd
import statsmodels.api as sm
import statsmodels.formula.api as smf

# Assuming you have a DataFrame 'df' with columns: y, subject, medianscore, attendance
# Convert 'subject' to categorical variable
df['subject'] = pd.Categorical(df['subject'])

# Fit the linear mixed-effects model
model_formula = 'y ~ medianscore + attendance + worstscorescore'
random_effects_formula = '0 ~ subject'

```

The right window shows a user message 'use pymc instead' and a response from ChatGPT: 'Certainly! Here's an example of how you can implement a linear mixed-effects model using PyMC3, a probabilistic programming library in Python:' followed by a code block:

```

python
import pymc3 as pm
import pandas as pd
import numpy as np

# Assuming you have a DataFrame 'df' with columns: y, subject, medianscore, attendance
# Convert 'subject' to a categorical variable
df['subject'] = pd.Categorical(df['subject']).codes

# Define the model
with pm.Model() as mixed_effect_model:
    # Fixed effects
    beta_0 = pm.Normal('beta_0', mu=0, sd=10)
    beta_medianscore = pm.Normal('beta_medianscore', mu=0, sd=10)
    beta_attendance = pm.Normal('beta_attendance', mu=0, sd=10)
    beta_worstscorescore = pm.Normal('beta_worstscorescore', mu=0, sd=10)

    # Create the mixed-effects model

```

Both windows have a 'Message ChatGPT...' button at the bottom. The footer of the slide reads 'Gen AI Assisted Data Science' and '61'.

61

Official (Open)

# Addressing Data Scarcity and Quality Issues

- Generative models can generate synthetic data when real data is limited.
- Implications for rare events or edge cases (anomaly)

The diagram illustrates an autoencoder architecture. An input image  $x$  (a green vertical bar with a '4' at the bottom) is processed by an 'Encoder' ( $g_\phi$ ) to produce a compressed latent representation  $z$  (a red square). This  $z$  passes through a 'Bottleneck!' (indicated by a grey arrow) and is then processed by a 'Decoder' ( $f_\theta$ ) to reconstruct the input as  $x'$  (a blue vertical bar with a '4' at the bottom). Above the diagram, text states 'Ideally they are identical.  $x \approx x'$ ' and 'An compressed low dimensional representation of the input.'

Source: <https://towardsdatascience.com/anomaly-detection-using-autoencoders-5b032178a1ea>

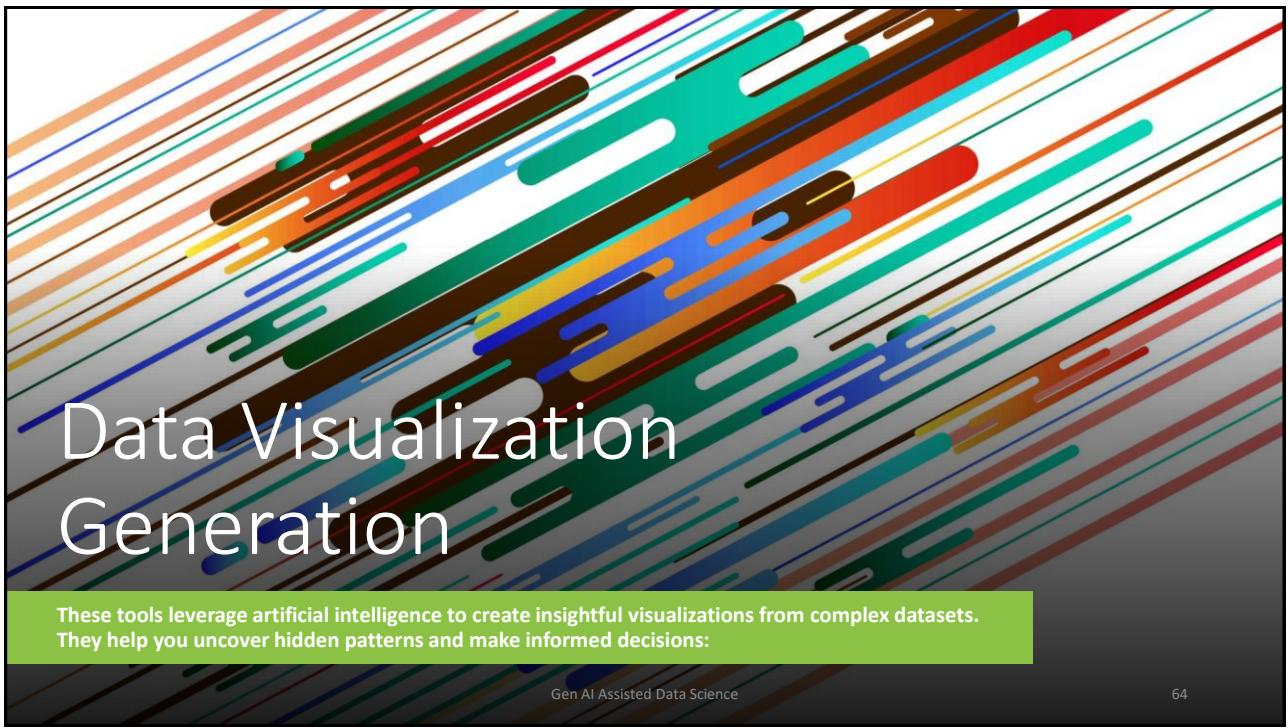
Gen AI Assisted Data Science

62



63

63



64

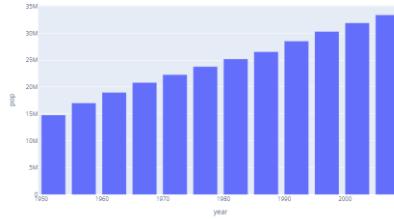
Gen AI Assisted Data Science

64

Official (Open)

## Manual Data Visualization Coding (Python Plotly)

```
import plotly.express as px
data_canada = px.data.gapminder().query("country == 'Canada'")
fig = px.bar(data_canada, x='year', y='pop')
fig.show()
```



Source: <https://plotly.com/python/bar-charts/#bar-chart-with-plotly-express>

Gen AI Assisted Data Science

65

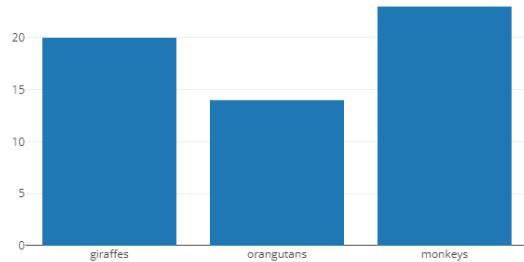
65

Official (Open)

## Manual Data Visualization Coding (JavaScript Plotly)

```
var data = [
{
  x: ['giraffes', 'orangutans', 'monkeys'],
  y: [20, 14, 23],
  type: 'bar'
};

Plotly.newPlot('myDiv', data);
```



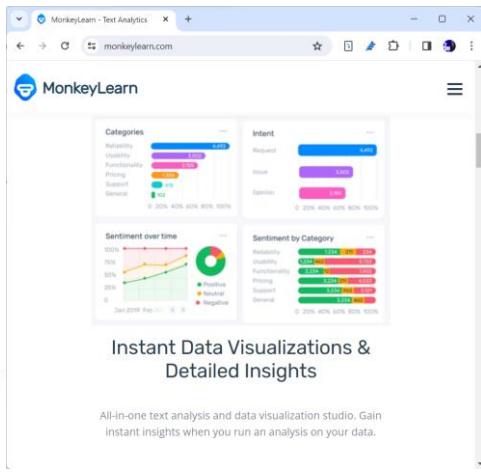
Source: <https://plotly.com/javascript/bar-charts/#basic-bar-chart>

Gen AI Assisted Data Science

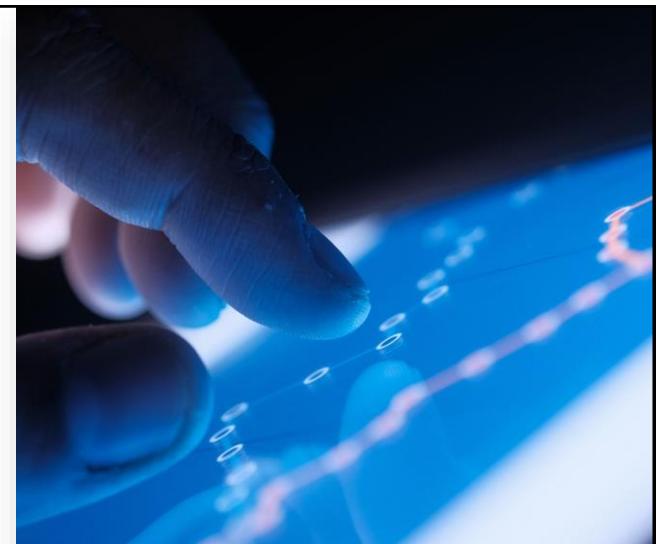
66

66

# MonkeyLearn



All-in-one text analysis and data visualization studio. Gain instant insights when you run an analysis on your data.



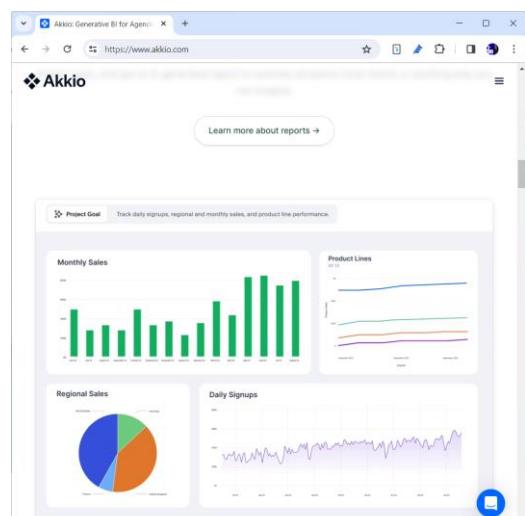
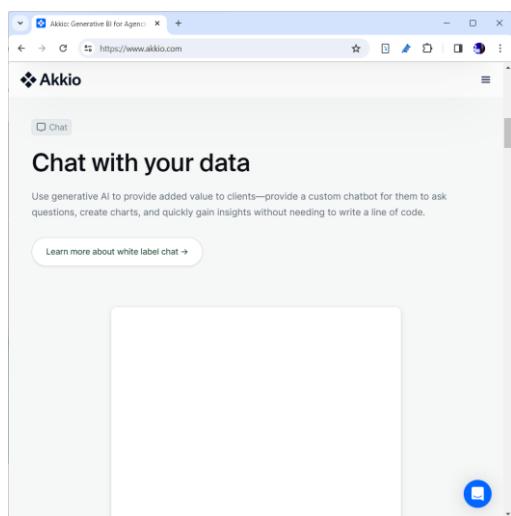
**A powerful data visualization platform that allows you to create interactive dashboards and reports**

Gen AI Assisted Data Science

67

67

# Akkio

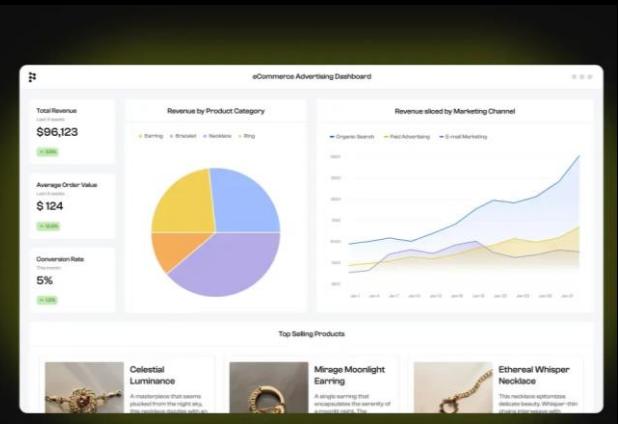
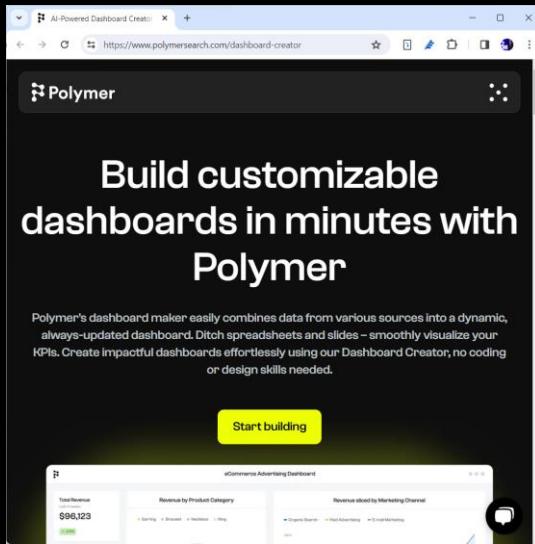


Gen AI Assisted Data Science

68

68

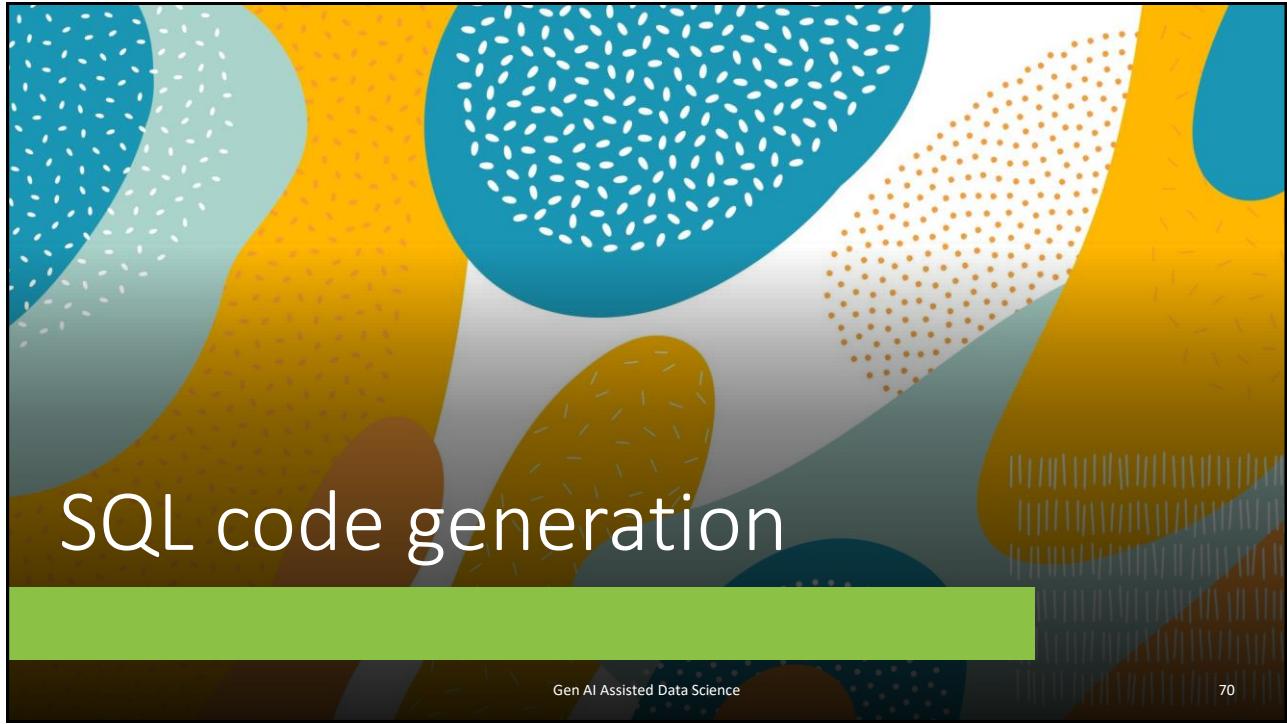
# Polymer



Gen AI Assisted Data Science

69

69



Gen AI Assisted Data Science

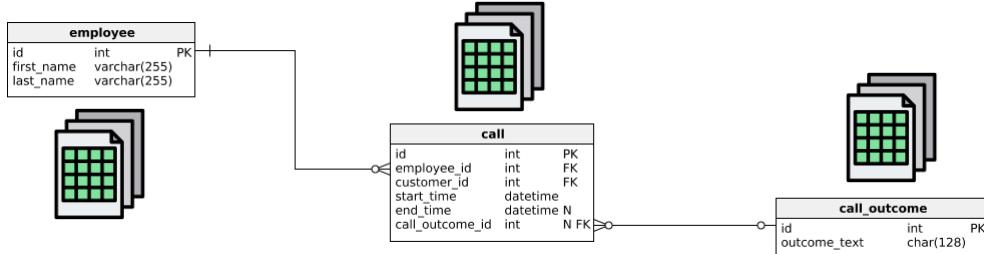
70

70

35

Official (Open)

## Manual SQL code (ANSI)



```

SELECT employee.first_name, employee.last_name, call.start_time, call.end_time, call_outcome.outcome_text
FROM employee
INNER JOIN call ON call.employee_id = employee.id
INNER JOIN call_outcome ON call.call_outcome_id = call_outcome.id
ORDER BY call.start_time ASC;
    
```

Gen AI Assisted Data Science

71

71

Official (Open)

## Manual SQL code (Oracle Variant)

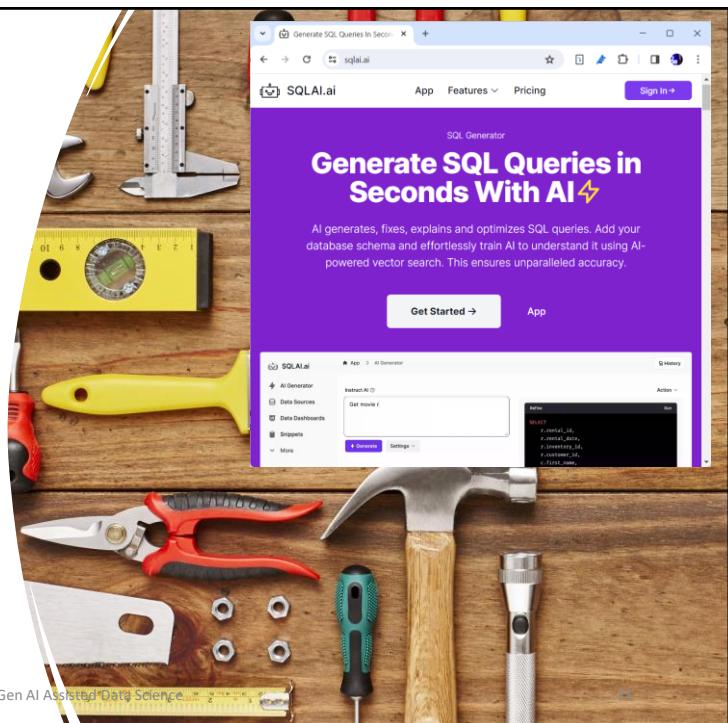
```

SELECT employee.first_name, employee.last_name, call.start_time, call.end_time, call_
outcome.outcome_text
FROM employee, call, call_outcome
WHERE call.employee_id = employee.id
AND call.call_outcome_id = call_outcome.id
ORDER BY call.start_time ASC;
    
```

Gen AI Assisted Data Science

72

72



**SQL AI**

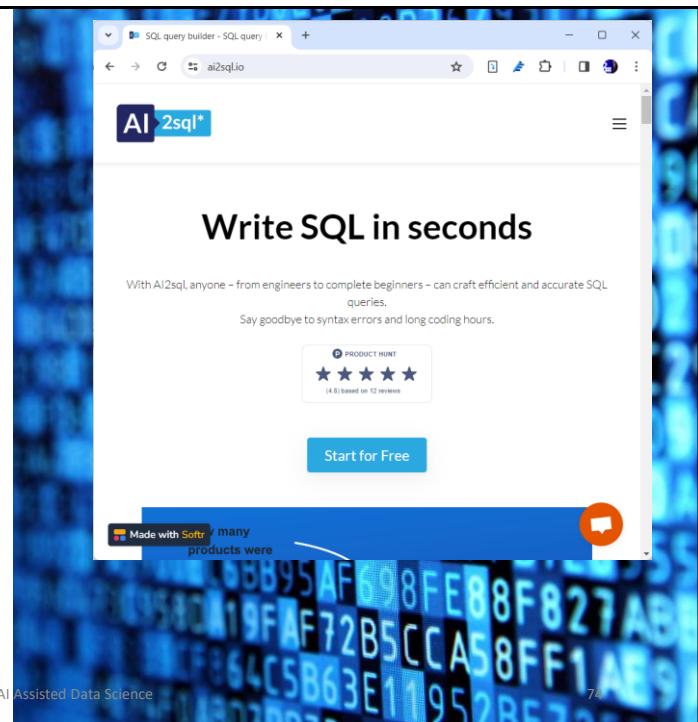
---

**Tools for Data Analysis and Engineering:**

**These tools enhance your SQL experience and streamline data engineering tasks**

The SQLAI.ai interface is shown in a browser window. The main heading is "Generate SQL Queries in Seconds With AI". Below it, a subtext states: "AI generates, fixes, explains and optimizes SQL queries. Add your database schema and effortlessly train AI to understand it using AI-powered vector search. This ensures unparalleled accuracy." A "Get Started →" button is visible.

73



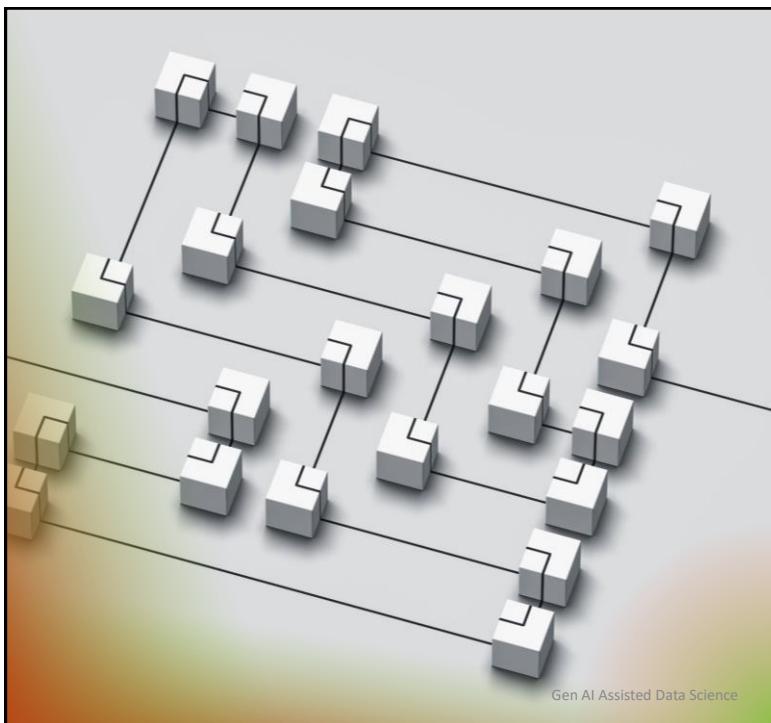
**AI2sql.io**

- Driven by AI, it generates SQL queries based on natural language instructions

The AI2sql.io interface is shown in a browser window. The main heading is "Write SQL in seconds". Below it, a subtext states: "With AI2sql, anyone – from engineers to complete beginners – can craft efficient and accurate SQL queries. Say goodbye to syntax errors and long coding hours." A "Start for Free" button is visible.

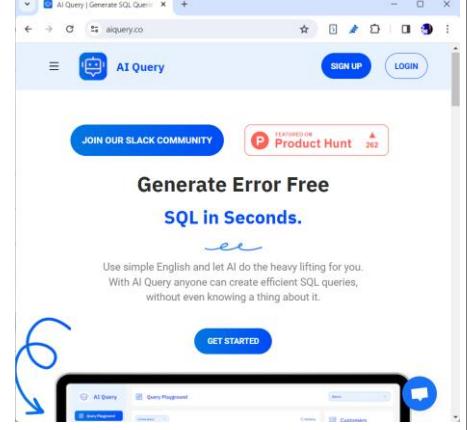
Made with Softy many products were

74



**AI Query**

- Utilizes the GPT-3 model to create SQL queries from plain language

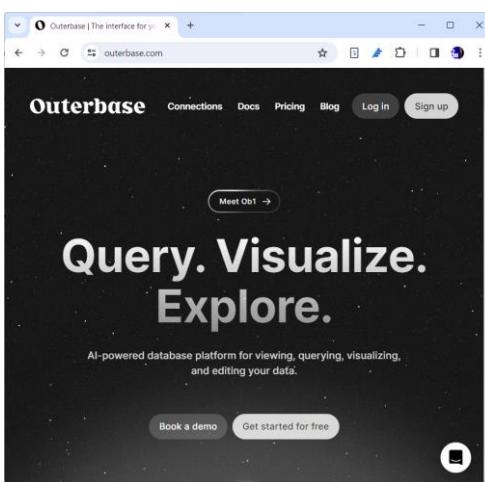


75

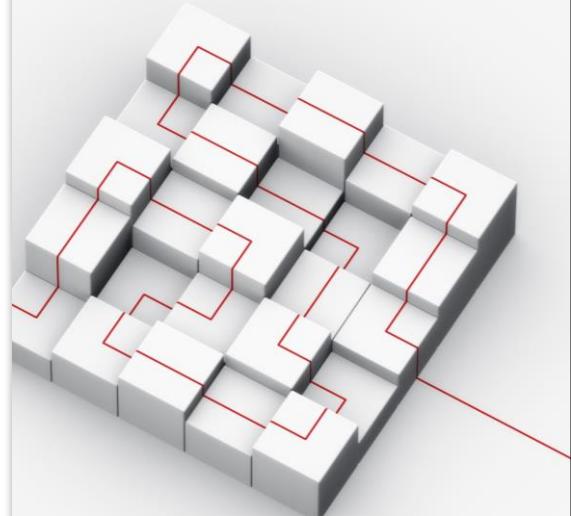
75

## Outerbase:

- Works on top of databases like Postgres, MySQL, and other relational databases



Gen AI Assisted Data Science



76

76

## SQL Chat

Gen AI Assisted Data Science

- A chatbot-style tool for SQL interactions

77

## Quills AI:

- **Purpose:** Simplify data analysis with seamless SQL and visual editing.
- **Features:**
  - **Natural Language Interaction:** Interact with your data using plain language, making exploration intuitive.
  - **Customized Insights:** Empower your team with tailored data insights

Gen AI Assisted Data Science

78

# Python code generation

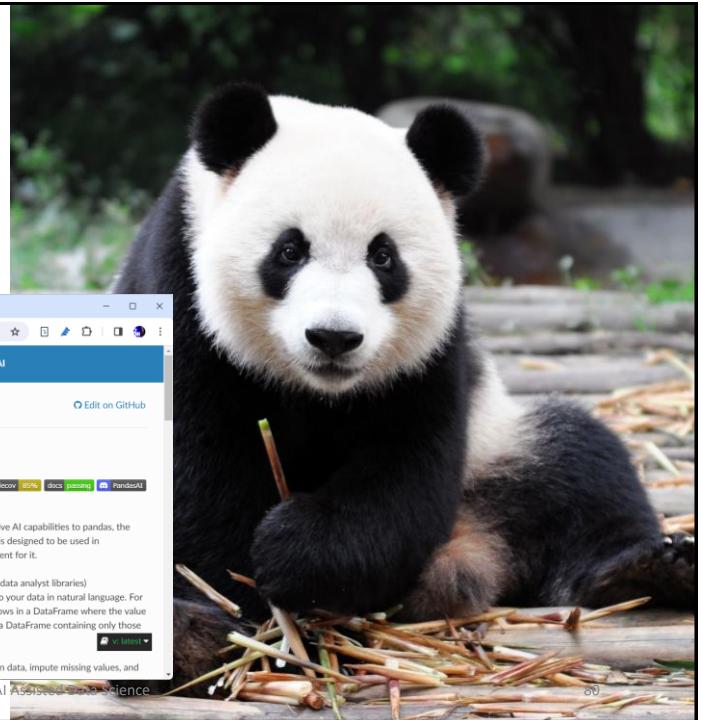
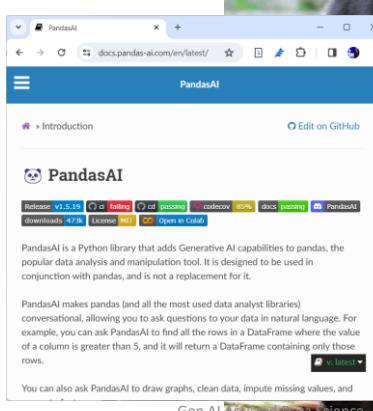
Gen AI Assisted Data Science

79

79

## Pandas AI:

- **Purpose:**
  - Generative AI-powered data analysis
- **Options:**
  - openai
  - azure-openai
  - google-palm



80

Official (Open)

The screenshot shows the official Amazon CodeWhisperer website. At the top, there's a navigation bar with links for Contact Us, Support, English, My Account, Sign In, and Create an AWS Account. Below the navigation, the page title is "Amazon CodeWhisperer" with a subtitle "Your AI-powered productivity tool for the IDE and command line". There are two main call-to-action buttons: "Use CodeWhisperer for free" and "Learn about CodeWhisperer customizations". To the right, there's a section titled "Start today with the Individual Tier:" listing benefits like "Free for individual use", "Unlimited code suggestions", "Reference tracking", and "50 security scans (per user, per month)". A "Learn more" link and a help icon are also present.

Amazon CodeWhisperer

Your AI-powered productivity tool for the IDE and command line

Use CodeWhisperer for free      Learn about CodeWhisperer customizations

Start today with the Individual Tier:

- Free for individual use
- Unlimited code suggestions
- Reference tracking
- 50 security scans (per user, per month)

Learn more     

Gen AI Assisted Data Science

81

ChatGPT Code Interpreter  
(Advanced Data Analysis)  
Plugin:

---

**Purpose:** Effective for data analysis tasks.

**Features:**

- Available under the ChatGPT Plus subscription.
- Web version currently available with no official IDE support.
- Designed to assist with data analysis tasks.

Gen AI Assisted Data Science

82

**GitHub Copilot.**

- Purpose: Accelerate development process and enhance productivity.
- Features:
  - Generates code suggestions based on context and prior code.
  - Supports multiple programming languages, including Python and SQL.
  - Integrates seamlessly with Visual Studio Code
  - Provides real-time auto-completion and code generation

83

83

**Pandas AI Demo**

Showcase a generative AI tool in action

84

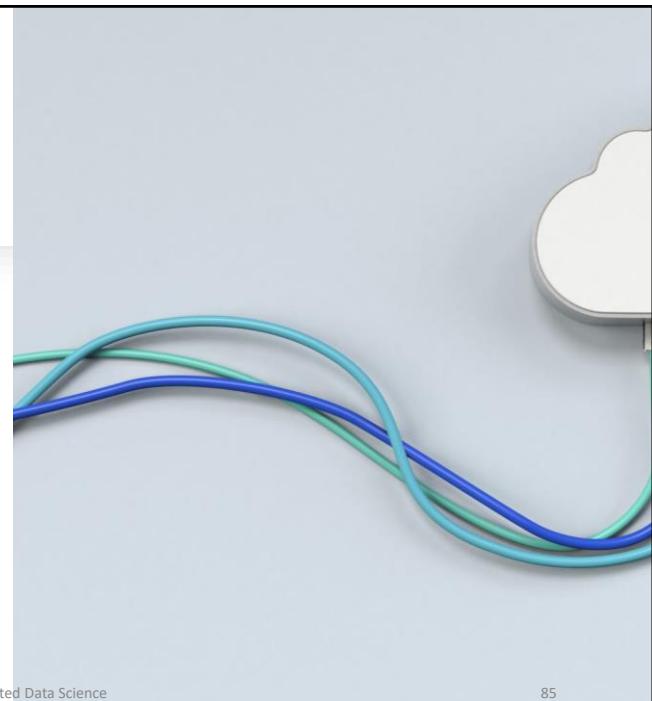
84

## Integration with Existing Workflows

- How generative AI fits into the data science pipeline.
- Integration considerations (APIs, libraries, cloud services).

Gen AI Assisted Data Science

85



85

## Fine-Tuning and Customization

- Adapting Pre-Trained Models
  - Transfer learning and fine-tuning.
- When to use pre-trained models vs. training from scratch.

Gen AI Assisted Data Science

86



86

Official (Open)

# Transfer learning vs. Fine-tuning

- **Transfer learning:**

- Uses pre-trained model weights as a starting point, **freezing them** to prevent modification.
- Adds new trainable layers specific to the new task.
- Best for **smaller datasets** and tasks with **similar features** to the pre-trained model.

- **Fine-tuning:**

- Also uses pre-trained model weights but **allows some layers to be updated** during training.
- Requires **more data** and closer task similarity for effective adaptation.
- Can achieve **better performance** on specific tasks compared to basic transfer learning.

Gen AI Assisted Data Science

87

87

Official (Open)

# When to use pre-trained models vs. training from scratch

## Use a pre-trained model when:

- **Limited data:** If you have a small dataset, leveraging the learned features from a pre-trained model can significantly boost performance compared to training from scratch on limited data.
- **Faster development:** Pre-trained models offer a "head start" on learning, saving you time and resources compared to building a model from the ground up.
- **Similar task:** If your task shares features and characteristics with the pre-trained model's task, it can adapt effectively with fine-tuning.
- **Computation limitations:** Training large models from scratch requires significant computational power. Pre-trained models often offer comparable performance with less computational investment.

## Train from scratch when:

- **Unique task:** If your task is very different from existing pre-trained models, building a specific model might be necessary for optimal performance.
- **Data access:** If you have access to a large, high-quality dataset specific to your task, training from scratch might be more effective than adapting a pre-trained model.
- **Interpretability:** In some cases, understanding the model's reasoning is crucial. Training from scratch allows for building models with more interpretable architectures.
- **Customization:** If your project requires complete control over the model's architecture and training process, building from scratch offers deeper customization.

Gen AI Assisted Data Science

88

88

44

Official (Open)

## Additional considerations

### Additional factors to consider:

- **Availability of pre-trained models:** Not all tasks have readily available pre-trained models with good performance.
- **Expertise required:** Training from scratch requires more expertise in machine learning compared to fine-tuning a pre-trained model.
- **Computational resources:** Both approaches require resources, but training from scratch often demands more.

Gen AI Assisted Data Science

89

89



## Balancing Fine-Tuning and Overfitting:

- **Trade-off between model performance and generalization.**
- **Tips for avoiding overfitting during fine-tuning.**
  - **Domain adaptation** — Domain adaptation is a concept in machine learning where the aim is to adapt the knowledge learned from one domain (source domain) to another domain (target domain).

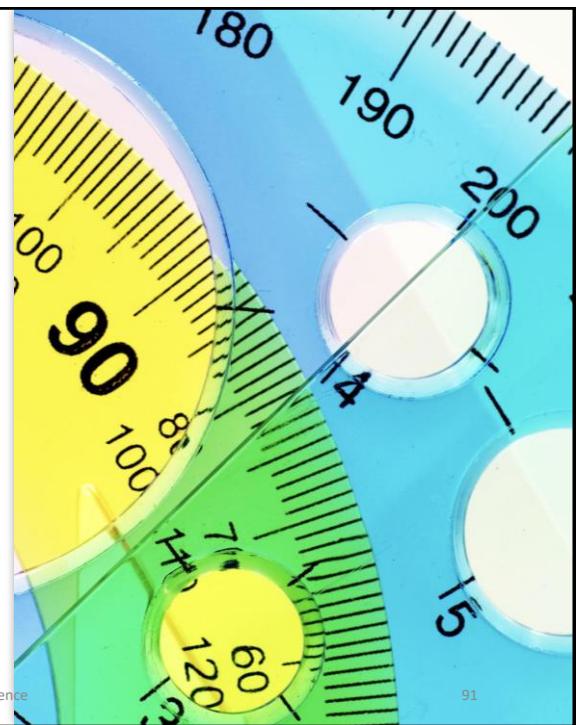
Gen AI Assisted Data Science

90

90

## Hyperparameter Tuning for Generative Models:

- Hyperparameters specific to generative tasks.
  - More focus on architecture choices (e.g. latent space dimensions, network depth)
  - Different domain specific metrics for Generative AI (e.g. FID)
  - Unstable training dynamics (e.g. mode collapse) make tuning tricky



91

## Collaboration Between AI and Data Scientists



Gen AI Assisted Data Science

92

92

## Synergy Between Human Expertise and AI Assistance



Gen AI Assisted Data Science

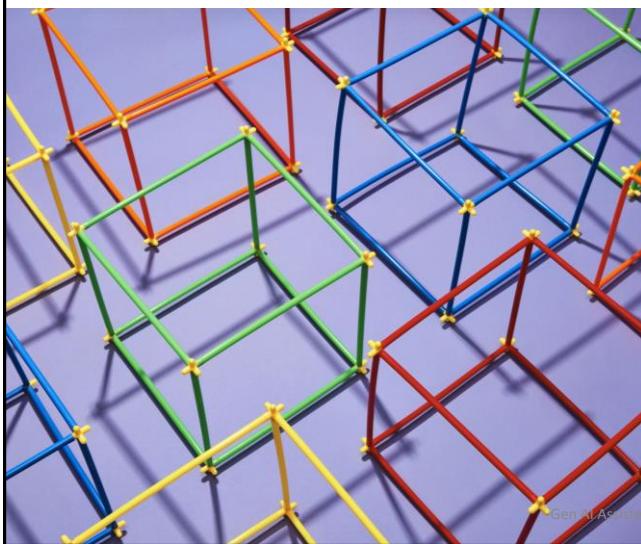


- Complementary roles of data scientists and generative AI.
- Collaboration rather than replacement.

93

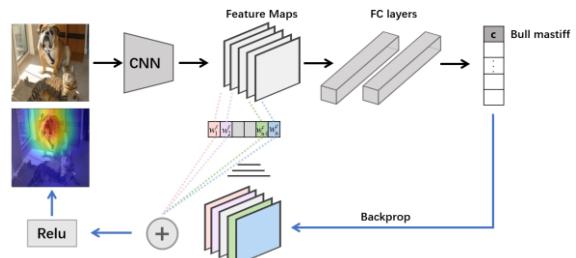
93

## Building Trust and Transparency



Gen AI Assisted Data Science

- Concerns about “black-box” models.
- Interpretability techniques (e.g., attention maps).

Source: <https://www.mdpi.com/2076-3417/12/8/3846>

94

94

## Ethical Considerations

- Biases in generative models.
- Responsible use of AI-generated content.



95

Official (Open)

Create a photo realistic picture of delicious steak dinner



Gen AI Assisted Data Science

96

96

Official (Open)

Create a photo realistic image of delicious hainanese chicken rice



Gen AI Assisted Data Science

97

97

Q&A and Closing  
Remarks

Gen AI Assisted Data Science

98

98

49

Official (Open)

## Audience Questions

- Q&A

Gen AI Assisted Data Science

99

99

Official (Open)

## Recap Key Takeaways

- AI is not new
- Generative AI build upon what narrow AI has provided
- Generative AI does not make exact copies, novel data is created
- Generative AI assists in the data science workflow
- Generative AI does not replace data scientists, but helps them to be more productive in their tasks

Gen AI Assisted Data Science

100

100

50

The End

---

[peterleong@sp.edu.sg](mailto:peterleong@sp.edu.sg)

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

Gen AI Assisted Data Science

101