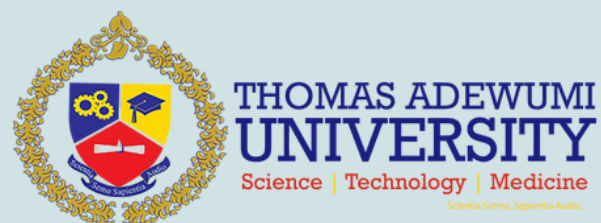


Open Source Data Science and AI

SWE 406

4th of July, 2025



OLUWATODUNNI OBAFEMI
22/10MSS013

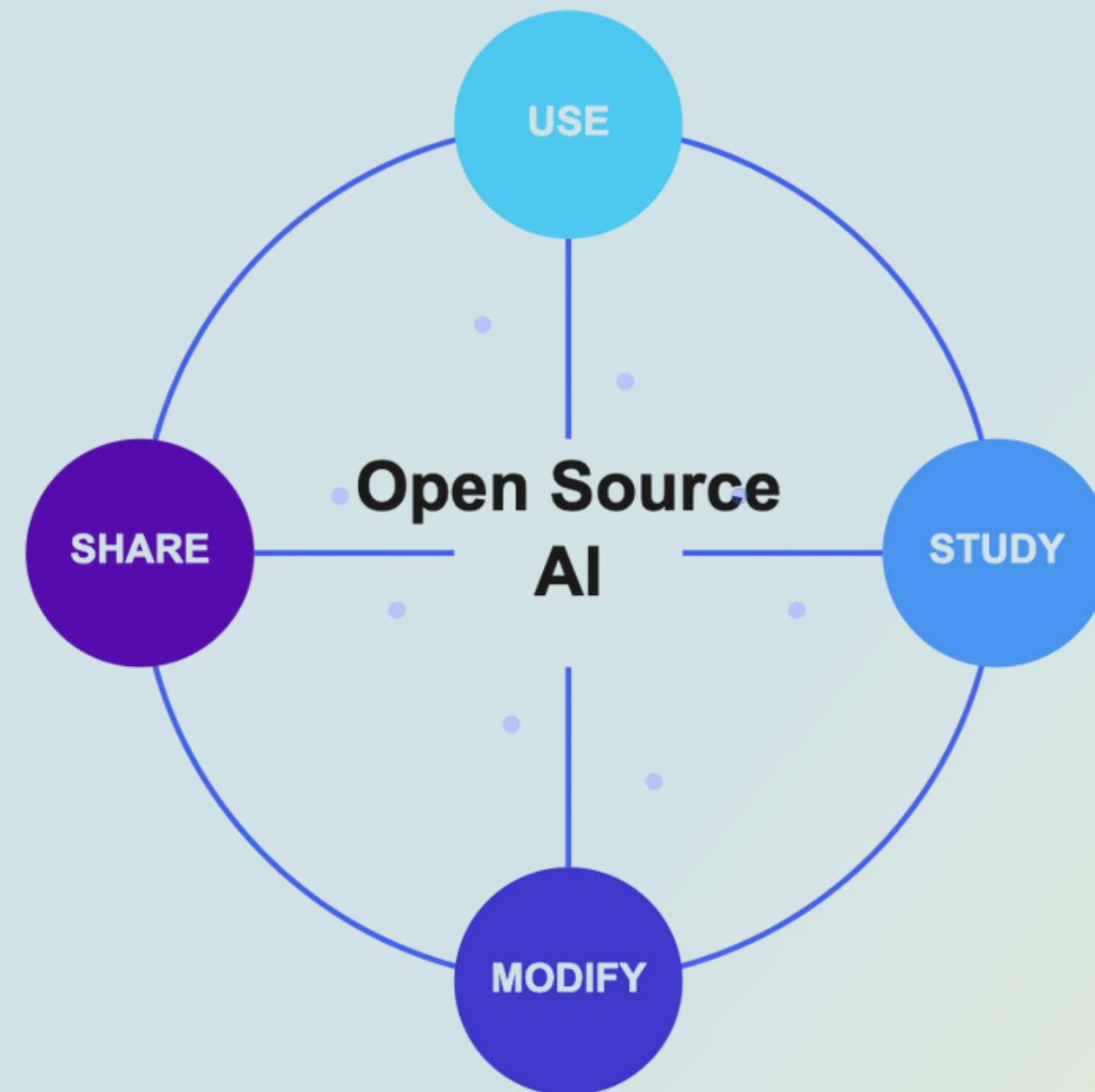
CONTENTS

- **Open-source tools for data science and AI (Python ecosystem, TensorFlow, PyTorch)**
- **Collaborative data science with open tools**
- **Open datasets and their applications**
- **Ethical considerations in open AI development**

SECTION ONE:

Open-Source Tools for Data Science and AI

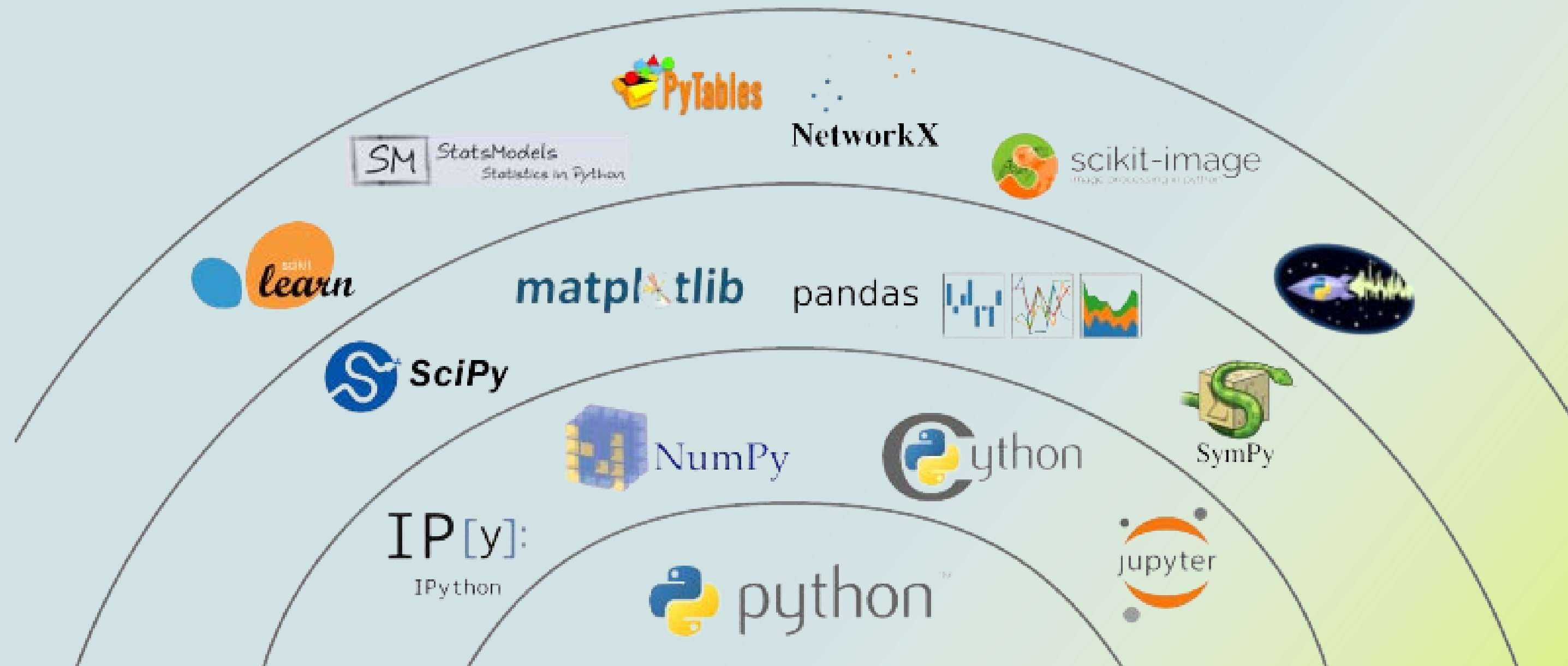
What is open source?



Why open source matters in AI?

- Fosters collaboration
- Accelerates innovation
- Democratizes access
- Enhances transparency and security

Python: The Heart of Open Data Science



- Easy-to-learn syntax & vast ecosystem
- Seamlessly integrates with AI frameworks
- Dominant language in AI and data science

Core Libraries for Data Science

NumPy

NUMERICAL COMPUTING

- An open source mathematical and scientific computing library for Python programming tasks.
- The name NumPy is shorthand for Numerical Python.
- The NumPy library offers a collection of high-level mathematical functions including support for multi-dimensional arrays, masked arrays and matrices.

Pandas

DATA MANIPULATION

- A Python library used for working with data sets.
- It has functions for analyzing, cleaning, exploring, and manipulating data.
- The name "Pandas" has a reference to both "Panel Data", and "Python Data Analysis" and was created by Wes McKinney in 2008.

Matplotlib / Seaborn

DATA VISUALIZATION

- Matplotlib is a library in Python that enables users to generate visualizations like histograms, scatter plots, bar charts, pie charts and much more.
- Seaborn is a visualization library that is built on top of Matplotlib. It provides data visualizations that are typically more aesthetic and statistically sophisticated.

Scikit-learn

MACHINE LEARNING ALGORITHMS

- Scikit-learn's metrics enable thorough evaluation of machine learning models across different tasks and scenarios.
- Understanding these metrics helps in interpreting model performance, identifying potential areas for improvement and ultimately selecting or optimizing the best-performing model for a specific problem.

TensorFlow vs PyTorch

BOTH ARE OPEN SOURCE



TensorFlow

Written in C++ and is, as a result, very fast and efficient.

Feature rich; TensorFlow can be used for training data as well as for inference.

Very good documentation; TensorFlow has many users and a big community which has led to strong documentation.

High popularity; TensorFlow has established itself as the most used ML library over a number of years now.

Many APIs available; TensorFlow is a library with a rich choice of easy to use APIs.

Supports JavaScript; TensorFlow supports JavaScript, C++ and Java in addition to Python.

For Mobile & IoT, inferences can be performed with TensorFlow Lite on mobile devices such as Android or iOS, as well as on Edge TPU or Raspberry Pi.

PyTorch

Written in Python making it more accessible and flattening the learning curve. However, the C++ core means PyTorch is still quite fast.

Very flexible; as data size can also be changed during data training.

Popular at research level; PyTorch was by far the most talked about ML library at CVPR, one of the most important computer vision conferences.

Rapid growth in popularity in both business and research use cases.

Many libraries available; PyTorch is composed of multiple libraries and platforms.

Python-based; PyTorch allows developers to write code in Python

PyTorch API; the PyTorch API is often preferred as it is better designed - plus TensorFlow has historically changed their API frequently.

Additional Open Source AI Tools

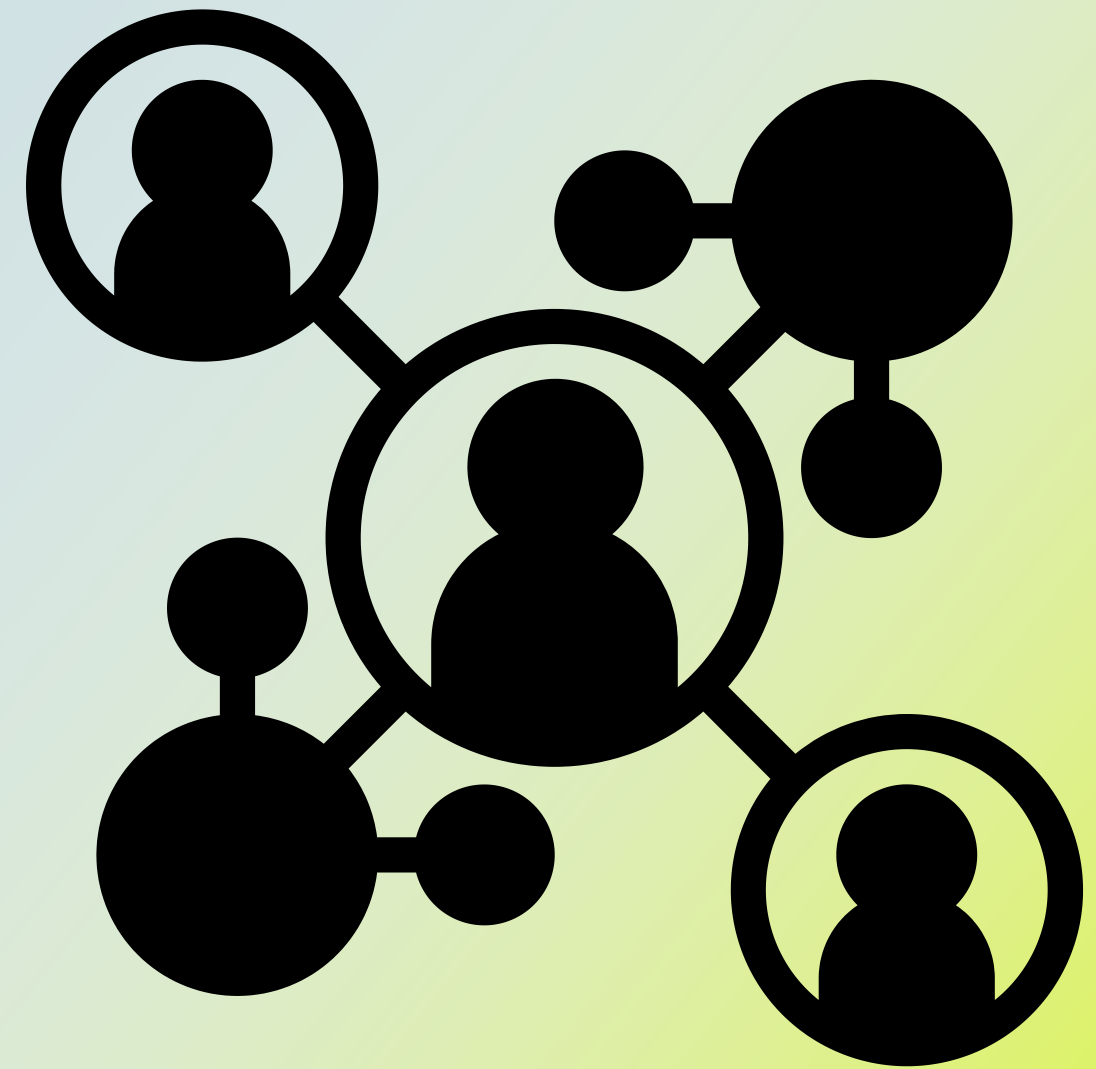
AI TOOL	USE
HuggingFace Transformers	Natural Language Processing: acts as the model-definition framework for state-of-the-art machine learning models in text, computer vision, audio, video, and multimodal model, for both inference and training.
Keras	High-level deep learning: Simplifies the process of building deep neural networks by providing a python interface for it
JAX	High Performance Machine Learning: a Python library for high-performance numerical computing, particularly suited for machine learning, that combines features of NumPy with automatic differentiation and just-in-time (JIT) compilation.

SECTION TWO:

Collaborative Data Science with Open Source Tools

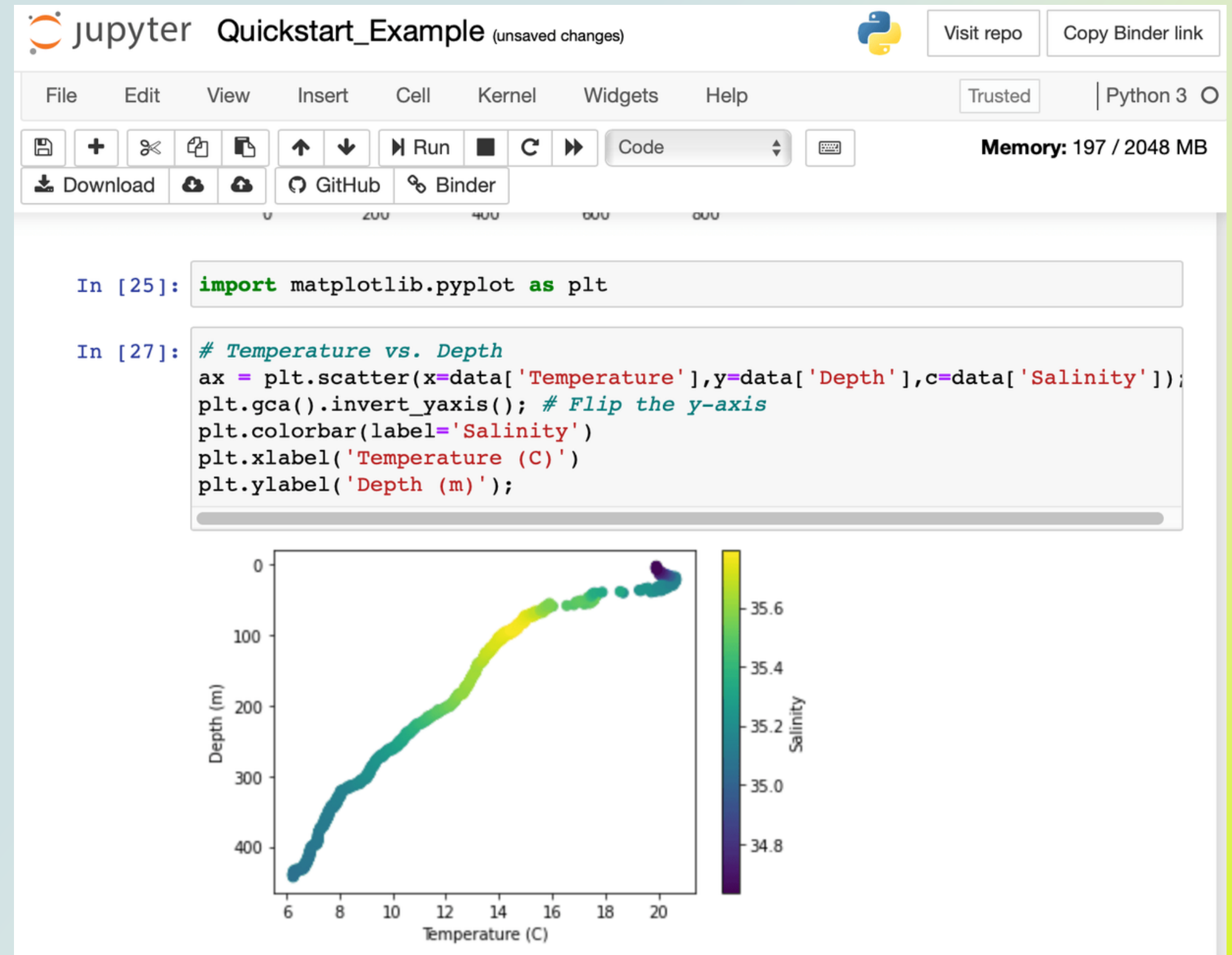
The Need for Collaboration in Data Science

- Data projects are multidisciplinary
- Collaboration boosts quality and innovation
- Open tools remove barriers



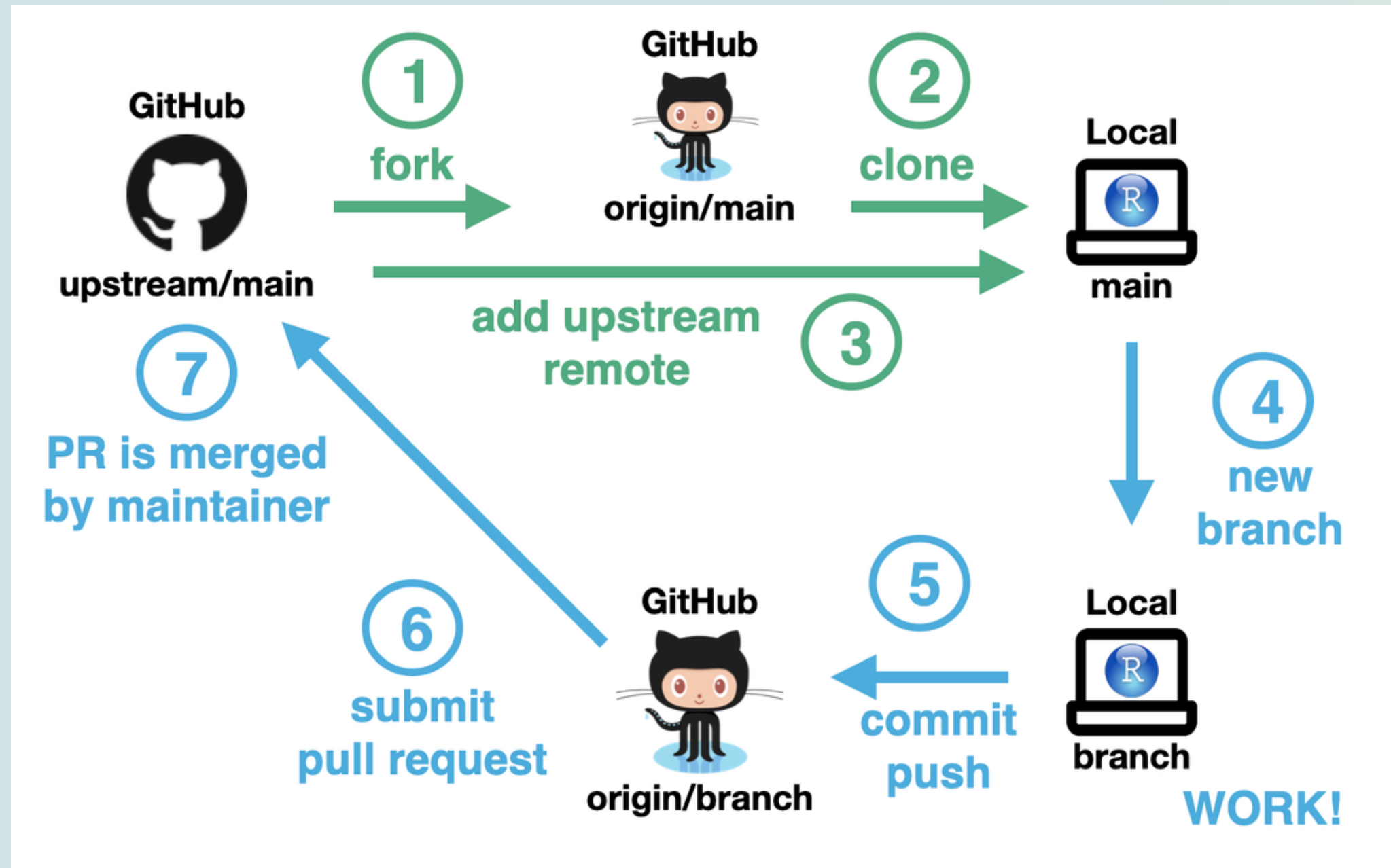
Jupyter Notebooks – A Data Scientist’s Best Friend

- Interactive coding & output in one place
- Ideal for exploration, documentation, and sharing
- Supported by JupyterHub for teams



Git & GitHub – Managing Code Together

- Version control to track changes
- Enables peer review and contribution
- Fork, clone, pull request = smooth collaboration



Other Platforms for Open Collaboration

- HuggingFace Hub
- Colab
- Kaggle
- BitBucket
- GitLab

kaggle



ATLASSIAN
Bitbucket



Google
colab



SECTION THREE:

OPEN DATASETS AND THEIR APPLICATIONS

WHAT ARE OPEN DATASETS?

- Freely accessible data for public use
- Used for research, benchmarking, experimentation
- Often backed by governments or organizations



Top Sources for Open Datasets

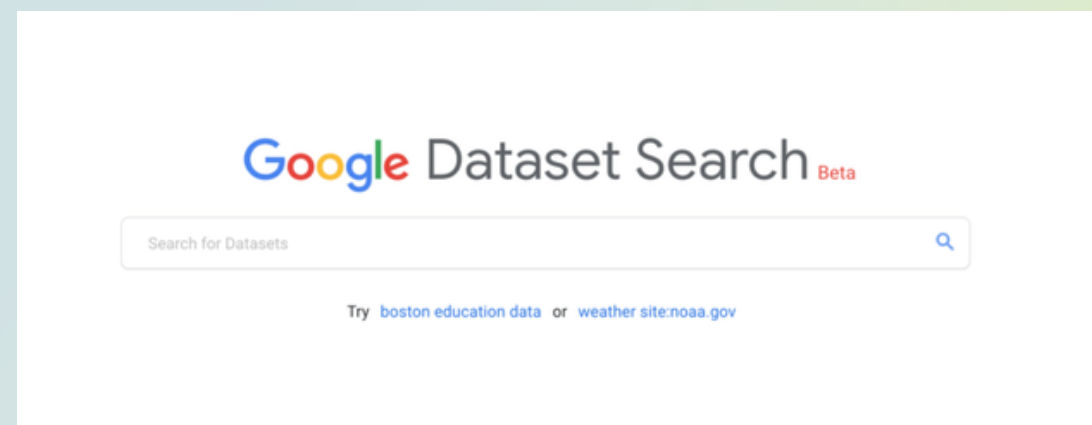
- Kaggle Datasets
- Google Dataset Search
- World Bank
- WHO
- UCI ML Repository
- OpenML
- Data.gov



kaggle



World Health
Organization



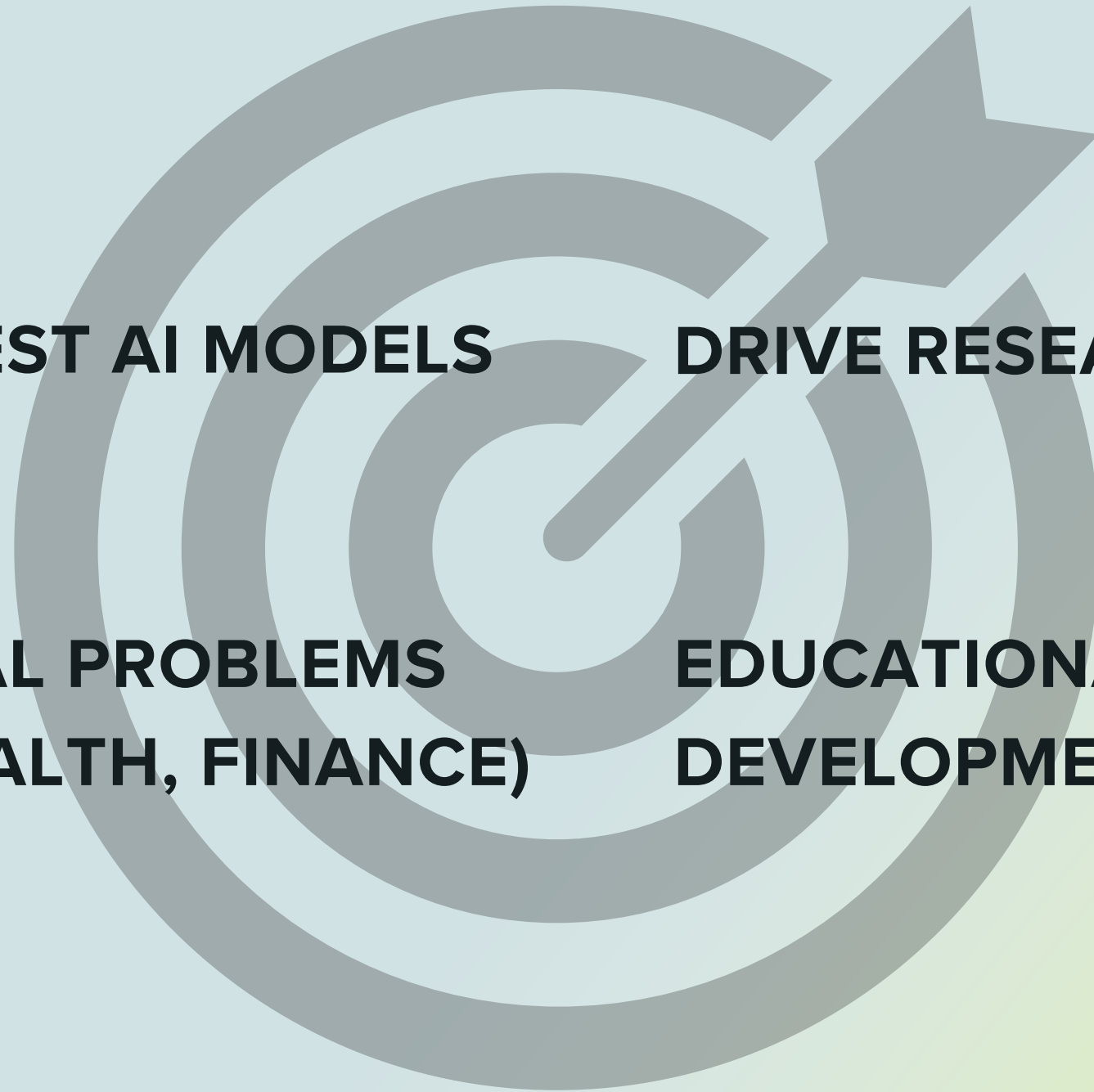
Applications of Open Datasets

TRAIN AND TEST AI MODELS

DRIVE RESEARCH AND POLICY

**SOLVE GLOBAL PROBLEMS
(CLIMATE, HEALTH, FINANCE)**

**EDUCATIONAL USE AND SKILL
DEVELOPMENT**



Example Case – ImageNet



- 14 million+ labeled images
- Pioneered computer vision breakthroughs
- Basis for models like ResNet, VGG, YOLO

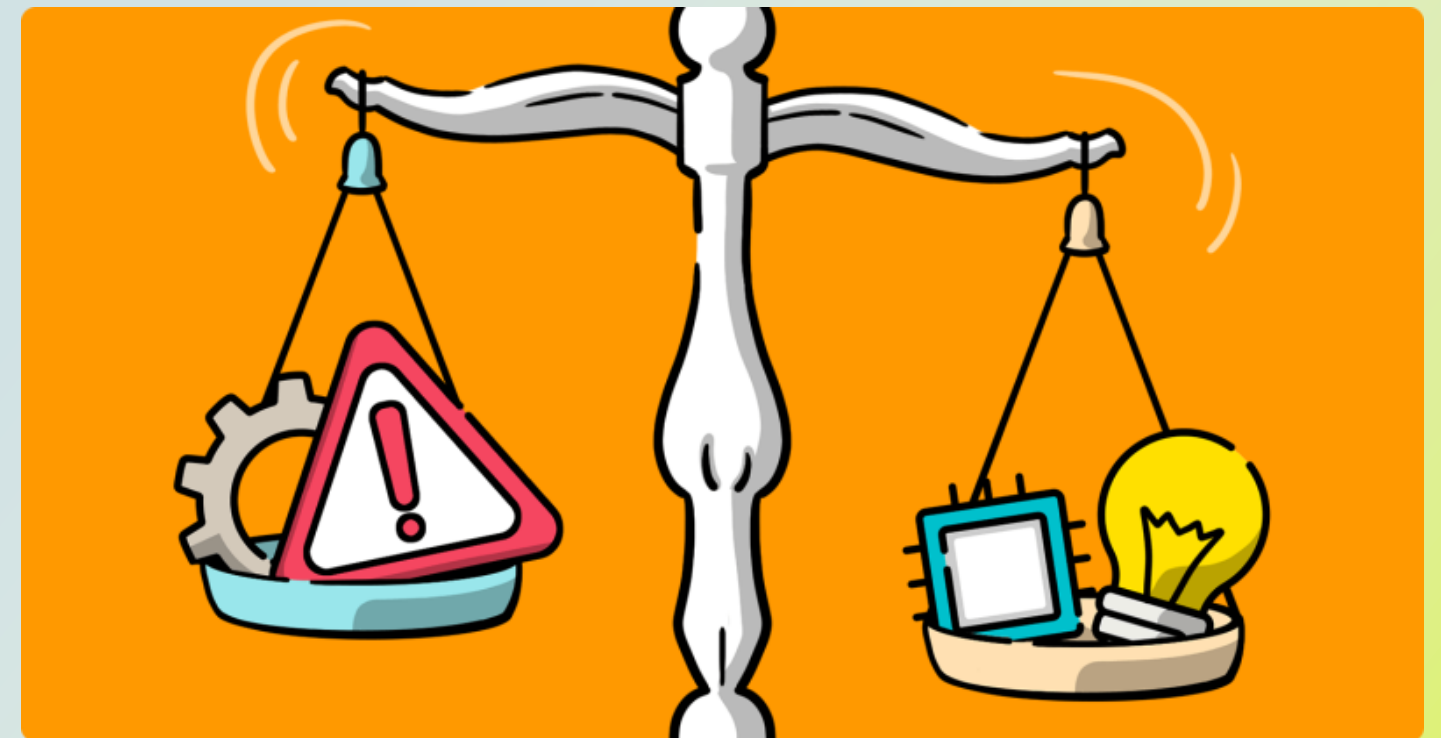
SECTION FOUR:

Ethical Considerations in Open AI Development

Why Ethics in Open AI?

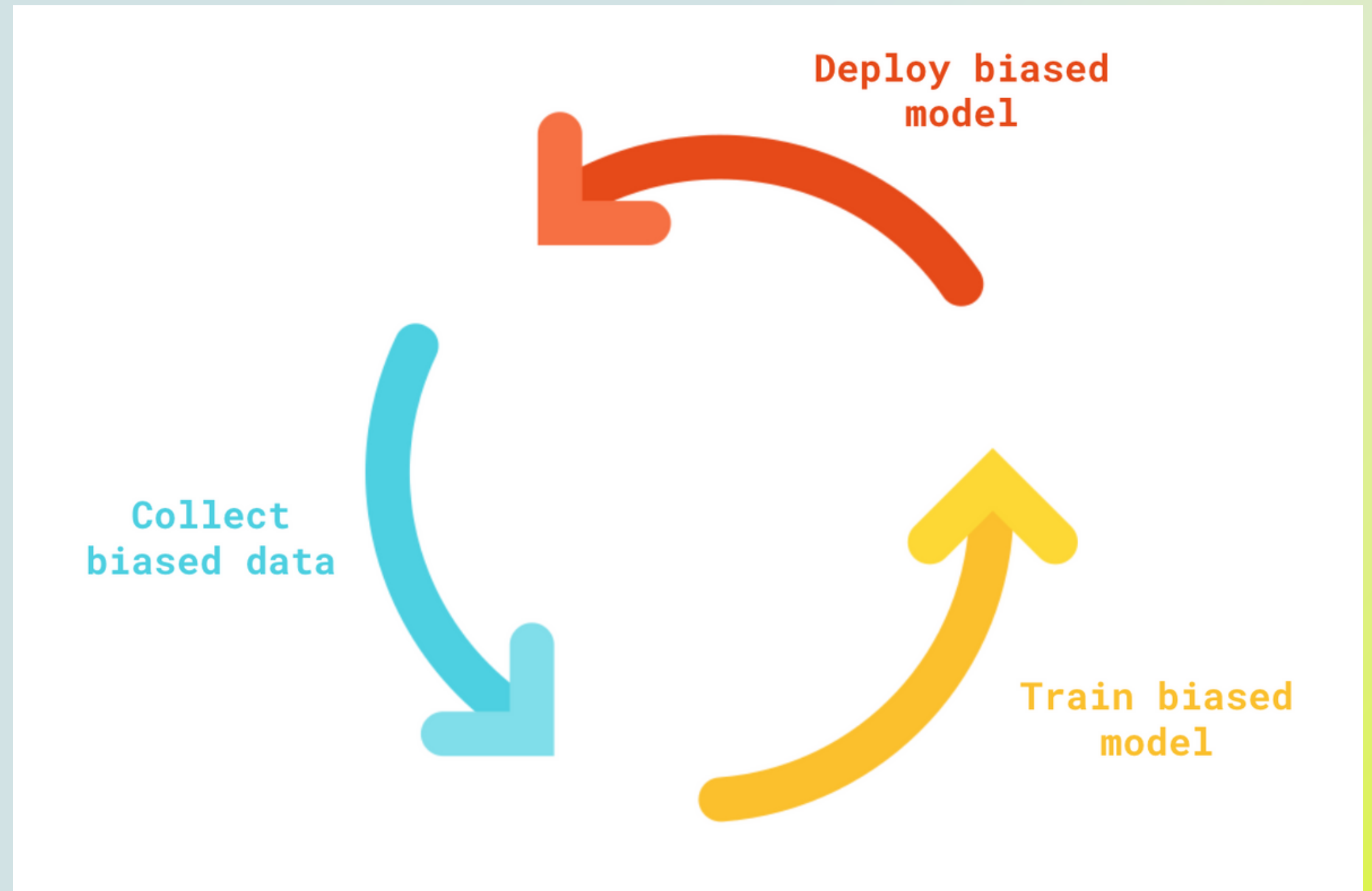
- AI can amplify bias and injustice
- Open source = more eyes = more accountability
- Need for transparency and safeguards

Responsibility vs Innovation



Bias in Open Datasets

- Historical data may reflect societal biases
- Models trained on biased data = biased predictions
- Need for diverse, representative datasets



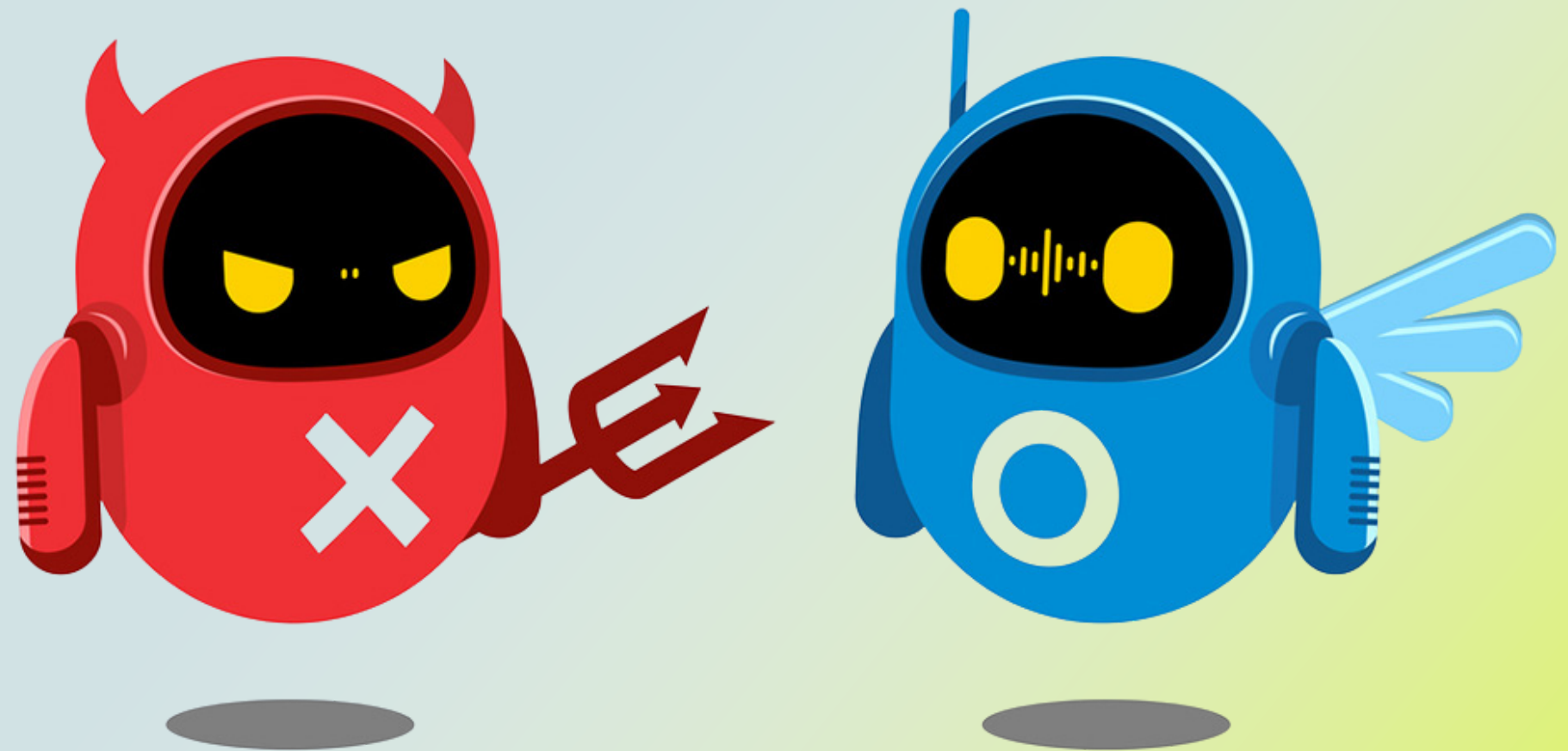
Transparency and Licensing

- Open-source = visible decisions & algorithms
- Respect attribution, privacy, and terms
- Choose licenses wisely (MIT, GPL, Apache)

License	Can be mixed with non-free software	Modifications can be taken private and not returned to author	Can be re-licensed by anyone	Contains special privileges for the original copyright holder over user's modifications	Restrictiveness
General Public License (GPL)	No	No	No	No	Strong reciprocity
GNU Library General Purpose License	Yes	No	No	No	Standard reciprocity
Berkeley System Distribution (BSD)	Yes	Yes	No	No	Permissive
Netscape Public License	Yes	Yes	No	Yes	Standard reciprocity
Mozilla Public License ²	Yes	Yes	No	No	Standard Reciprocity
Public Domain ³	Yes	Yes	Yes	No	Permissive (Not OSS)

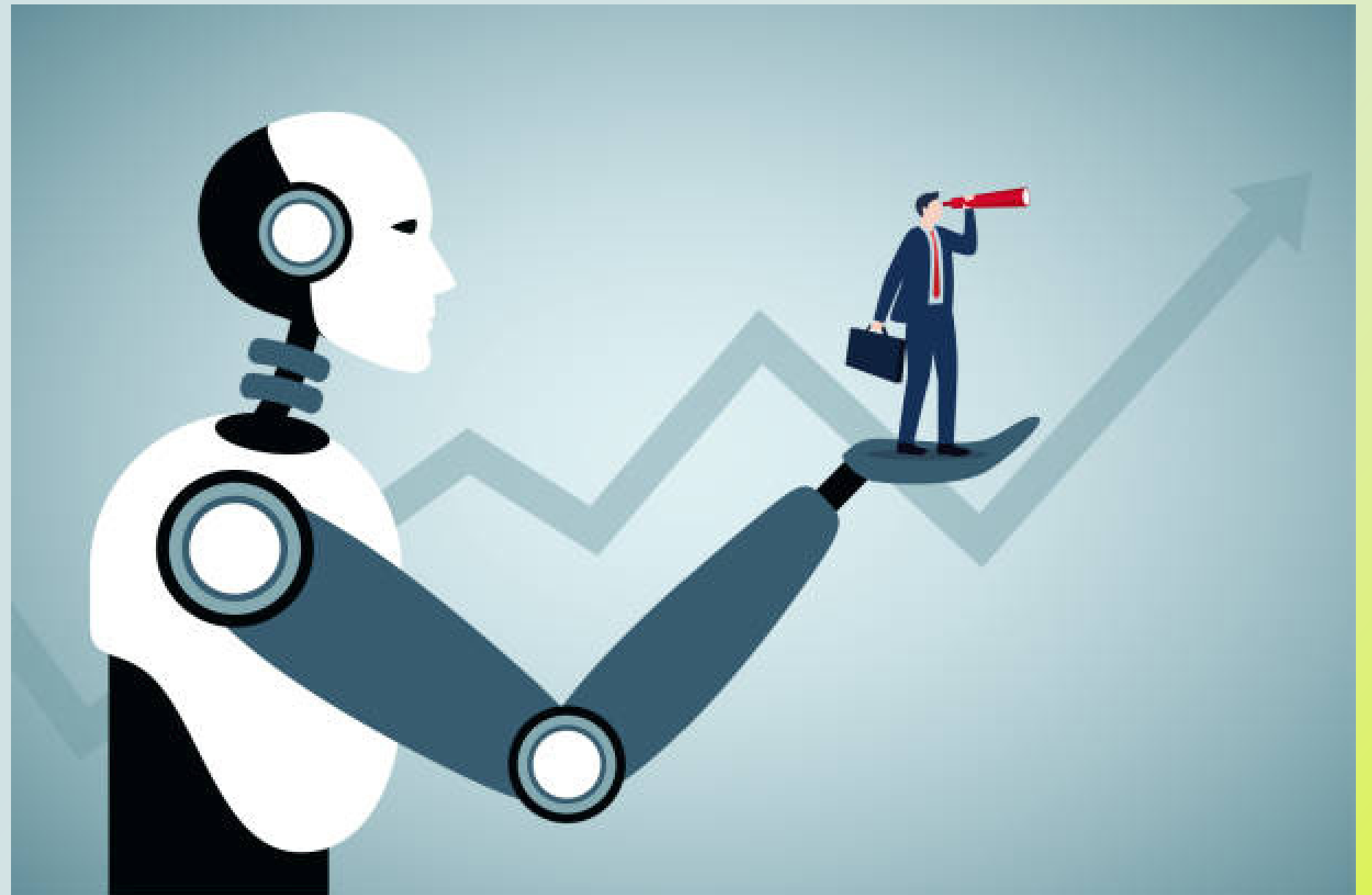
Preventing Misuse of Open AI

- Dual-use dilemma: Good vs Harm
- Add documentation & use cases
- Consider ethical review before releas



Future of Open-Source AI

- Global participation and innovation
- Ethical frameworks must grow alongside code
- Empowering the next generation of data scientists



CONCLUSION

Utilization

OPEN TOOLS + OPEN DATA

Collaboration

Collaboration is the new
superpower

Responsibility

Build AI responsibly, build it
together



**THANK
YOU**





QUESTIONS?