From Data to Discovery

A Baseline on
AI, Machine Learning & Pipelines
for Scientists



AI is Not Magic, It's Pattern Recognition

AI systems learn rules from examples rather than being explicitly programmed with those rules.

Traditional programming: Scientist defines the exact rule

Machine Learning (ML): Scientist provides examples (inputs + known outputs), and the computer finds the rule.

Why this matters for you: Many biological and physical phenomena are "easy to spot but hard to describe" formally . ML excels where intuition outpaces explicit formulation.

Analogy: *Teaching a student to identify a bird by showing many photos, rather than writing a textbook definition of every feather shape.*



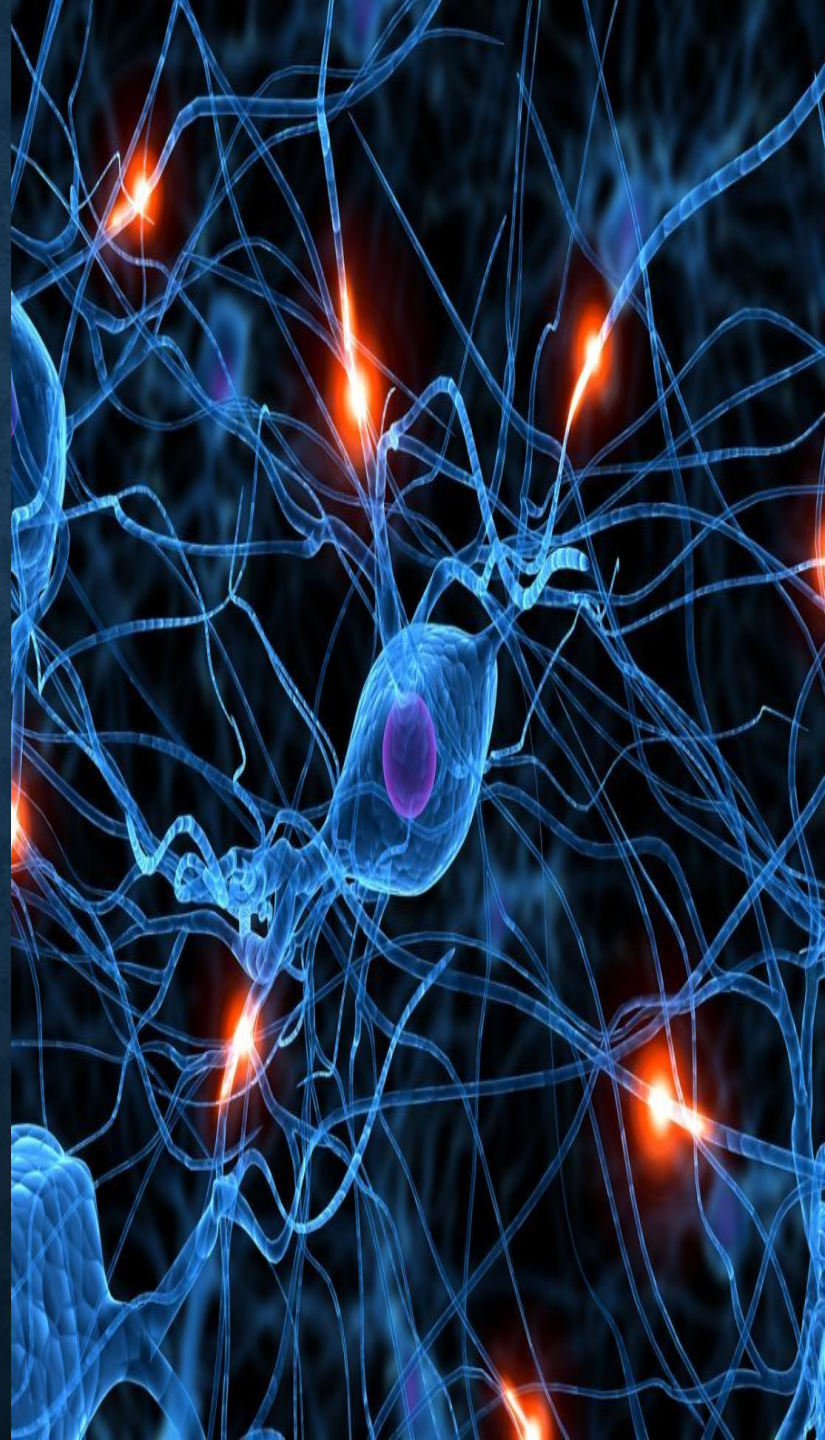
- **Artificial Intelligence (AI):** The broad goal of making computers mimic intelligent behavior.
- **Machine Learning (ML):** The subset of AI where systems learn from data (e.g., classification, regression, clustering).
- **Deep Learning (DL):** A specific ML technique using multi-layered neural networks. It excels at finding complex patterns in raw, unstructured data (images, sequences, spectra)

Deep Learning requires lots of labeled examples (thousands to millions). It is not always the right tool for small datasets

Deep learning builds complex concepts from simple building blocks.

- **The Perceptron:** A mathematical unit that takes inputs, applies weights (importance), sums them, and passes the result through an activation function.
- **Hierarchy of Features:** In image analysis, early layers detect edges → middle layers detect shapes/parts → deep layers detect whole objects (e.g., a cell nucleus)

Relevance to Science: *This hierarchy is why deep learning is powerful for microscopy, spectroscopy, and genomic sequences—it learns the features you might not know to look for.*



The Workflow

ML Pipelines Are Just Experimental Protocols

1. ***Data Acquisition/Integration***: **Like gathering samples from multiple freezers or collaborators.**
2. ***Exploration & Cleaning:*** **Like checking for contamination, outliers, or missing metadata.**
3. ***Feature Engineering:*** **Like deciding what to measure (e.g., converting raw spectra to normalized peaks).**
4. ***Modeling (Training):*** ***Like running the assay on a calibration set.***
5. ***Evaluation (Testing):*** **Like validating on a blinded sample set.**
6. ***Deployment/Monitoring:*** **Like freezing a stable protocol and periodically recalibrating .**

Why pipelines matter: **Reproducibility. If you change one step, you can rerun the whole chain without starting from scratch.**

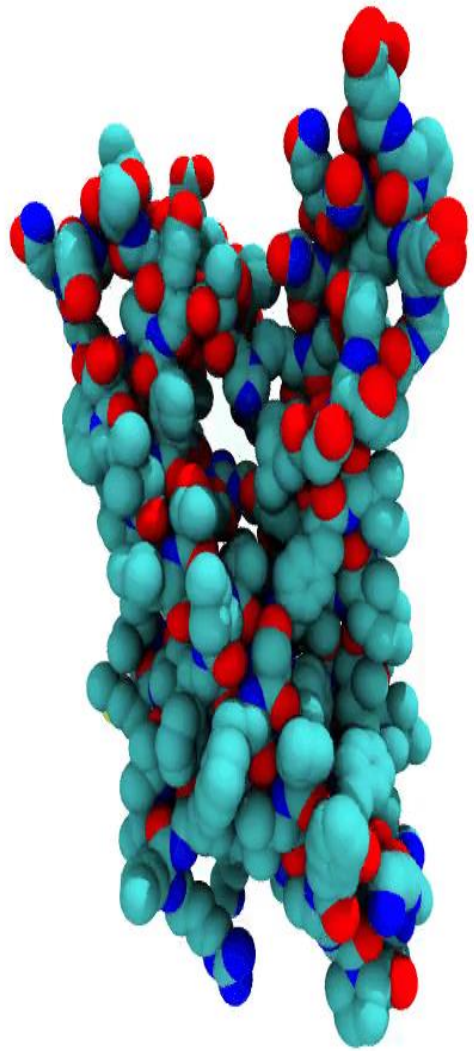
Applied Example 1
Life Sciences (Protein Discovery)

The Problem: **Microalgae proteins are mixed with bacterial noise; traditional tools miss them or take weeks.**

The ML Solution: Trained a model to read protein sequences like a language model reads text.

The Result: **Near-perfect separation of algal proteins from background noise, operating 10,000× faster than existing methods**

Pipeline connection: **This is a classification task on sequence data—a form of feature extraction that replaces manual annotation.**



Applied Example 2
Life Sciences (Drug Repurposing)

The Problem: **Finding a molecule that reverses disease-associated gene expression requires screening millions of chemicals.**

The ML Solution: A neural network trained on millions of chemical-gene expression measurements to predict how a compound's *structure* affects gene expression.

The Result: **Identified compounds for aggressive liver cancer (HCC) and idiopathic pulmonary fibrosis (IPF). Two new HCC compounds reduced tumors in mice .**

Pipeline connection: **Uses supervised learning (structure → expression profile) to prioritize candidates before wet-lab validation.**



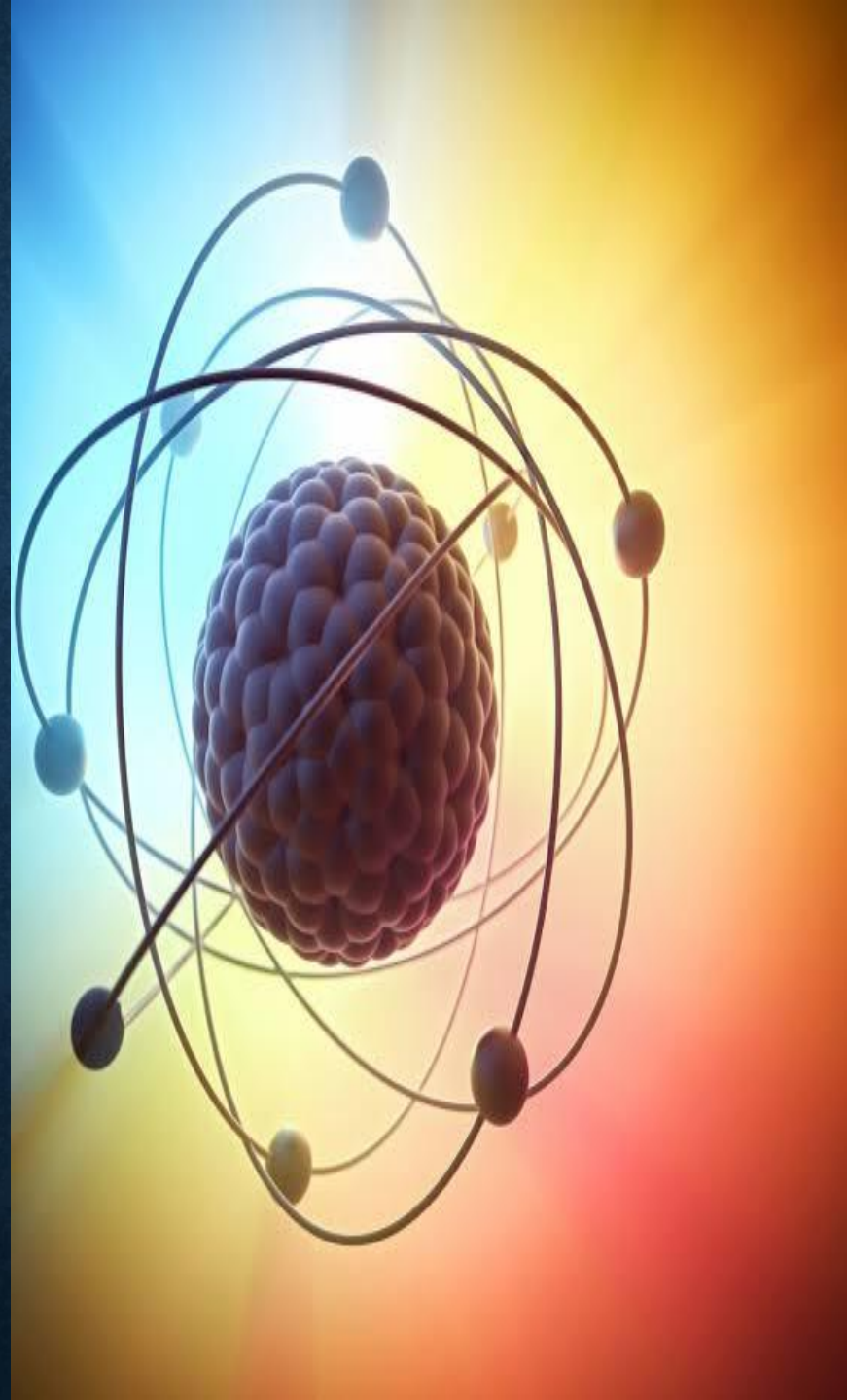
Applied Example 3
Physical Sciences (Theory Design)

The Problem: **Neutrinos have mass, but the Standard Model doesn't explain why. Searching the space of possible theories is vast.**

The ML Solution: : Reinforcement Learning—the AI proposes theoretical physics models, evaluates them against data, and improves its search strategy over time.

The Result: **Reproduced known neutrino theories and identified *new* candidate models in unexplored mathematical frameworks .**

Pipeline connection: **This is not prediction from data; it is AI-assisted hypothesis generation—a new paradigm for theory-heavy fields.**



The Universal Pipeline
(Cross-Disciplinary View)

The pipeline stages are identical; only the data type changes

Pipeline Stage	Biology (Protein)	Medicine (Drug)	Physics (Neutrino)
Input	Amino acid sequences	Chemical structures (SMILES)	Symmetry groups & particle data
Representation	Tokenized sequence	Molecular graph	Mathematical parameters
Task Type	Classification	Regression/Classification	Reinforcement Learning (Search)
Output	Protein identity	Gene expression prediction	Candidate theory models

Key Takeaways for Your Research

1. ***AI is a tool, not a replacement.*** **It handles scale (millions of compounds, petabytes of sequence) and pattern detection in high-dimensional data.**

2. ***Data quality > Model complexity.*** **Biological data is noisy. A clean, well-labeled dataset with a simple model beats a messy dataset with a deep network .**

3. ***You don't need to build the pipeline from scratch.*** **Many stages (cleaning, splitting, training) are standardized and available in open-source libraries or institutional platforms .**

4. ***Collaboration is the new normal*****. The MSU drug discovery project involved over 20 researchers across computer science, bench biology, chemistry, and clinical medicine .**

Prof. Mohamed El-Sayed Ahmed Abdel Aziz

Professor of Artificial Intelligence at Ajman University, UAE, and a researcher affiliated with the Artificial Intelligence Research Center (AIRC

- **His research interests include artificial intelligence, machine learning, deep learning, evolutionary computation, optimization, AI agents, and AI applications in healthcare, education, and smart systems. •**
- **He has published extensively in leading international journals and conferences and has led numerous research projects and collaborations with regional and international academic institutions.**
- **His recent work focuses on AI-driven decision support systems, intelligent agents, educational technologies, healthcare informatics, and responsible AI.**
- **He is actively involved in supervising graduate students, organizing scientific conferences, and contributing to initiatives that promote AI innovation, open science, and capacity building in the Arab region.**

Prof. Salah Obayya

Acting Executive Vice President for Research and Founding Director, Center for Photonics and Smart Materials, Zewail City of Science, Technology and Innovation, Egypt

---

- **UNESCO Chair-Holder, Paris, France**
- **Honorary Chair Professor of Photonics, University of Nottingham, UK**
- **He earned his doctorate from City, University of London,**
- **In 2002, he became an Associate Professor at Brunel University,**
- **In 2006, professorship at the University of Leeds.**
- **In 2008, a Chair Professor at the University of South Wales and founded the Photonics and Broadband Communications (PBC) Research Center, there, he also directed postgraduate programs**
- **In 2012, he returned to Egypt to join Zewail City of Science and Technology as a Chair Professor and Founding Director of the Center for Photonics and Smart Materials (CPSM).**
- **he took on roles as Vice Chairman of ZC and currently serves as the Director-General of Research Institutes.**
- **Editor-in-Chief of Optical and Quantum Electronics (2023).**
- **He also acted as Reviewer for many international Photonics Journals.**
- **He has received several national and international awards and recognitions for his scientific contributions.**


Thank
You