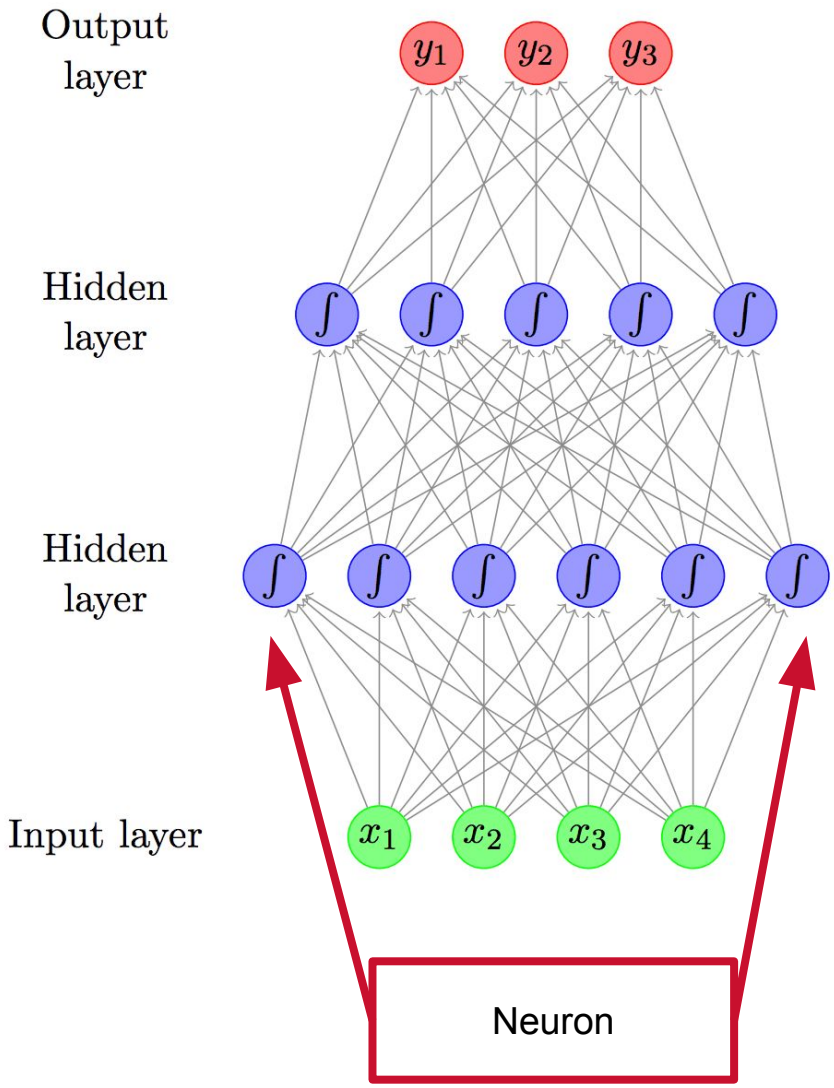
Explainable AI For Source Code Applications

SDMAY25-30

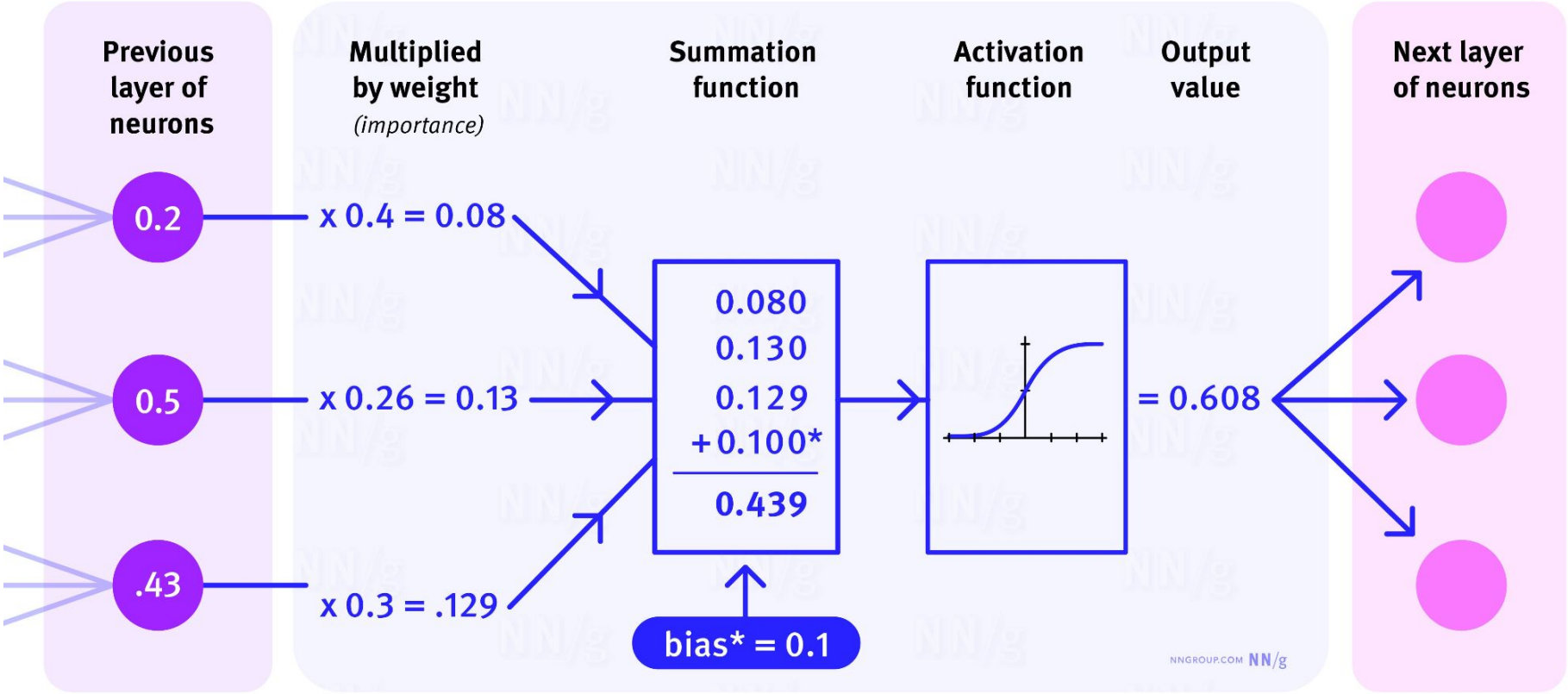
Manjul Balayar, Kellan Bouwman, Sam Frost, Akhilesh Nevatia, Ethan Rogers
Client: Dr. Ali Jannesari/ISU SwAPP Lab
Advisor: Arushi Sharma

IOWA STATE UNIVERSITY

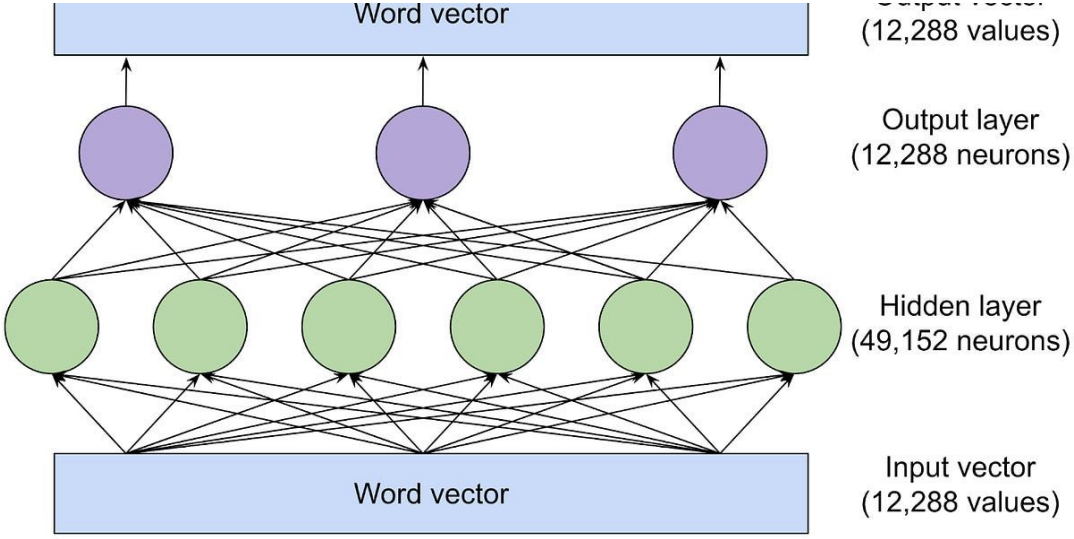
- Machine Learning / Statistical Learning:
  - Machine learning: A broad discipline that focuses on how computers can learn from data
  - Statistical learning: A branch of artificial intelligence that focuses on turning raw data into actionable information. Statistical learning theory is a framework that uses statistical and functional analysis to build models that can make predictions and draw conclusions from data
- Neuron:
  - A collection of a set of inputs, a set of weights, and an activation function.
  - It translates these inputs into a single output.
- Deep Learning:
  - Type of machine learning based on artificial neural networks in which multiple layers of processing are used to extract progressively higher level features from data

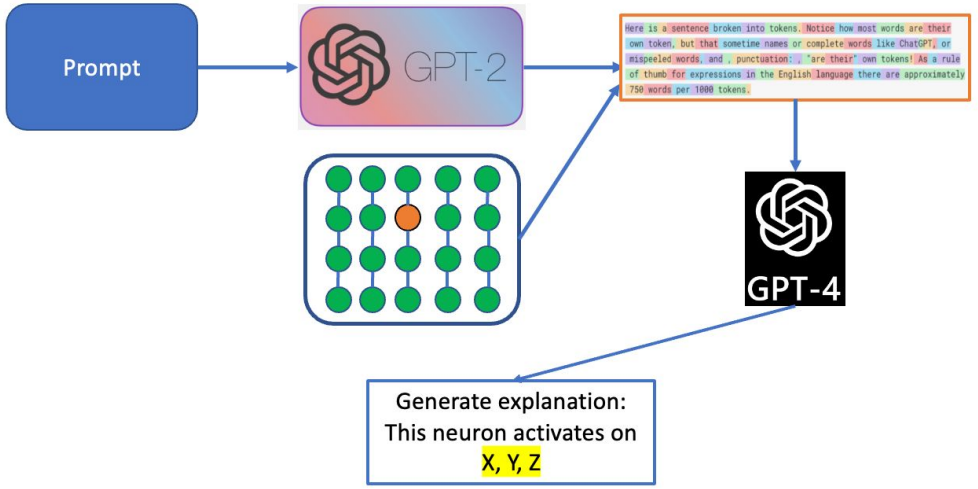


How a Single Artificial Neuron Works



- LLM:
  - Large Language Model
  - Form of Deep learning on NLP (Natural Language Processing)
- Generative AI:
  - A type of AI that uses generative models to create new content, such as text, images, videos, music, and audio
  - Based off an LLM (example: ChatGPT3.5)
- Neuron Activation
  - Non-linear function that we apply over the input data coming to a particular neuron and the output from the function will be sent to the neurons present in the next layer as input
  - A Path of Neurons





Project Overview

- Client
  - Dr. Ali Jannesari/ISU SwAPP Lab
- Abstract
  - Focus on auto-labeling code datasets using AST tools, regular expressions, and LLM-generated labels.
- Goal
  - Deepen the understanding of LLMs and generative AI by analyzing neuron activations and applying metrics and heuristics. The project will also unify two existing code bases into a flexible and scalable framework that can be deployed seamlessly across Colab, local environments, and HPC clusters.

Project Overview - Continued

- Users
  - Researchers
    - ML Engineers / Researchers
    - Prompt Engineers / Researchers
    - Computer Scientists
  - Students
    - Graduate Students
    - Undergraduate Students
  - Industry Professionals
- We aim to include students, researchers, and industry professionals as key users. Our documentation will prioritize accessibility and clarity for all experience levels, while the codebase will remain flexible and scalable to meet both academic and industry needs.

IDEALS Professional Responsibility - Doing Well

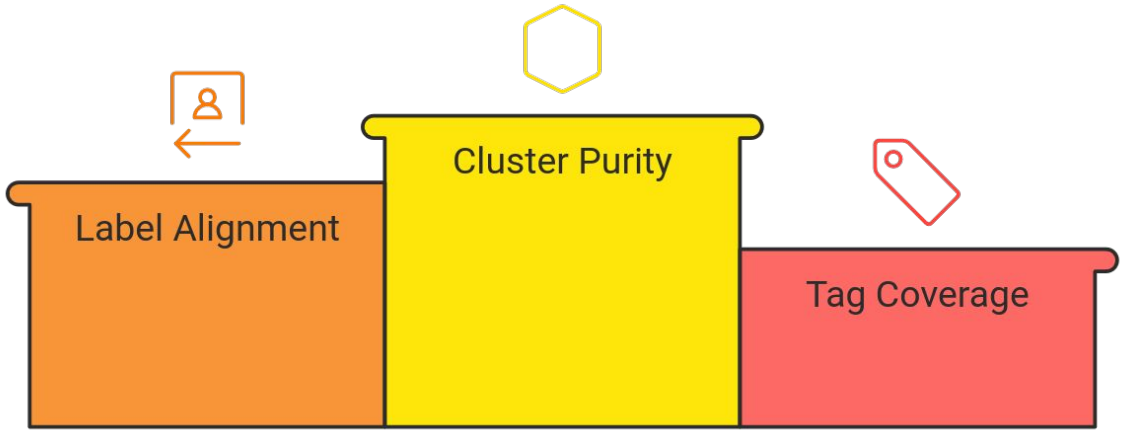
- Area: Rule/Duty-Based Ethics
- Relevance to Our Project:
  - Ensures compliance with industry codes and standards in software engineering.
  - Upholds data privacy and intellectual property rights when handling code datasets.
- Team's Approach:
  - Strict adherence to relevant IEEE and ISO standards throughout development.
  - Implementation of policies respecting licensing and usage rights of all code sources.
  - Regular ethical training and discussions to keep team members informed.
- Why This Upholds Ethical and Professional Responsibilities:
  - Demonstrates commitment to professional codes of conduct and ethical guidelines.
  - Protects the rights of individuals and organizations associated with the code.
  - Builds trust with stakeholders by operating transparently and responsibly.

IDEALS Professional Responsibility - Can Improve

- Area: Consequentialism
- Relevance to Our Project:
  - Understanding the broader impact and potential consequences of our AI tools on society.
- Team's Current Approach:
  - Focused mainly on technical development and meeting project milestones.
  - Less emphasis on assessing long-term societal implications and user impacts.
- How We Will Change to Better Uphold Ethical Standards:
  - Impact Assessments:
  - Incorporate evaluations of potential consequences into our planning process.
- Stakeholder Engagement:
  - Gather feedback from users and experts on possible societal effects.
- Mitigation Strategies:
  - Adjust project goals to reduce negative outcomes and enhance positive impacts.

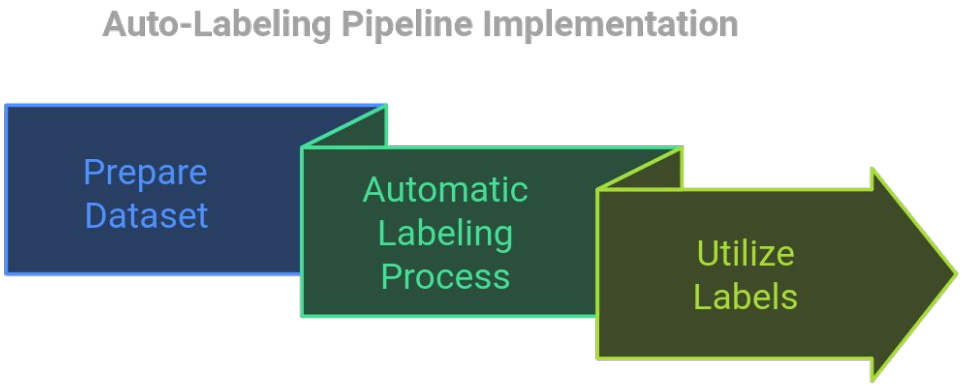
Broader Context Area-Four Principles Chart

- Highlighted Area: Privacy and Data Protection
- Importance to Our Project:
  - Handling code datasets that may contain sensitive or proprietary information.
  - Necessity to anonymize data and implement secure storage solutions.
  - Aligns with legal regulations and ethical standards in data handling.
- Actions Taken:
  - Established data anonymization protocols.
  - Restricted access controls for sensitive information.
  - Regular audits to ensure compliance with privacy laws.



Potential Ethical Issues We Are Concerned About

- Bias in AI Models:
  - Risk of perpetuating existing biases present in training data.
  - Potential unfairness in how code properties are interpreted and clustered.
- Misuse of Our Tools:
  - Auto-labeling pipeline could be exploited to infringe on intellectual property rights.
  - Need to promote responsible and ethical usage among users.
- Transparency and Explainability:
  - Challenges in making complex AI models understandable to all users.
  - Importance of providing clear explanations to avoid misinterpretation.



Virtue Important to Our Team

- Virtue: Collaboration
- Significance:
  - Values teamwork and open communication among team members.
  - Encourages sharing of ideas and constructive feedback.
  - Leads to a more cohesive and ethically sound project outcome.
- Implementation:
  - Regular team meetings and collaborative work sessions.
  - Inclusive decision-making processes.
  - Supportive environment fostering professional growth.