



Young AI, ML, Data Science, CS Research Innovators

From Public Data to Published Research

Week 1: Orientation & Introduction

Icebreaker

Name? Grade?

What word or phrase most captures the sense of wonder or potential you see in AI/ML?

Program Overview

What is CrunchTutor? What is Young Innovator AI,
ML, Data Science, CS Research?

What is CrunchTutor?



is an educational organization founded to promote **academic enrichment** and **student success**.

We aid in:

1. Teaching **educational concepts** (tutoring)
2. Guiding students to develop **strong study habits**
3. Enabling students to succeed outside of the classroom (**college applications, internship, academic competitions, etc.**)

What is Young Innovator AI, ML, Data Science, CS Research?

1. Enable students to gain **AI/ ML research experience**
2. Different project pathways (conference paper, application, web project) based on interests.
3. Help students create works that look good on **college applications, resumes, etc.**

Meet Your Coach - Ronel Solomon

- **MHS Alumni**
- PostGrad Data Science in USF
- **CrunchTutor Tutor/Coach**
 - Data Science
 - Python
 -

Email: ronelsolomon@gmail.com

WhatsApp: +1 (510) 861-8211



Session Logistics

- 90-minute live sessions for interactive learning.
- 60 minutes of offline/web-based building help weekly.
- Adjustments for longer sessions during summer.



CS Research

What do we do with it? Why is it important?

Program Goals

What will you complete within this program?

Foundations, Research Exploration & Coding Basics (Weeks 1-4)

1. Introduce the research process with a focus on AI/ML and data science.
2. Provide an optional coding bootcamp if new to programming.

Proposal Development & Data Familiarization (Weeks 5-8)

1. Develop a clear, testable research proposal focused on AI/ML.
2. Familiarize you with data manipulation, visualization, and introductory machine learning techniques.

Data Analysis & Research Report Drafting (Weeks 9-16)

1. Deepen data analysis skills using publicly available datasets.
2. Begin drafting a comprehensive research report with an AI/ML focus.

Writing, Revision & Final Presentation (Weeks 17-22)

1. Finalize the research report with rigorous editing and proper citation.
2. Prepare and practice presenting research findings.

Timeline

How will this program be structured?

Weeks 1-4

Phase 1.

**Foundations, Research
Exploration, & Coding
Basics**

Week 1: Orientation & Introduction

1. Overview of AI/ML research and the importance of data-driven studies.
2. Introduction to the program structure and goals.

Offline Help: Discuss personal interests in AI and Q&A.
Optional Coding Bootcamp (if needed): Basics of Python/Excel for data handling.

Week 2: LR Basics & Topic Brainstorming

1. How to search academic databases (e.g., IEEE Xplore, arXiv, Google Scholar) and evaluate research papers.
2. Identify gaps in current AI/ML research to select a literature review topic.
3. Developing a research question and formulating a hypothesis.

Offline Help: Hands-on review of selected articles and discussion on emerging trends.

Optional Coding Bootcamp: Introduction to programming fundamentals (data types, control structures).

Week 3: Intro to Public Data & Editing LRs

1. Overview of public data sources relevant to AI/ML (Kaggle, UCI Machine Learning Repository, Open Data initiatives) and data ethics.

Offline Help: Workshop on assessing data quality and initiating data exploration.

Optional Coding Bootcamp: Simple exercises in data import, cleaning, and initial exploratory data analysis.

Week 4: Drafting a Proposal

1. Identifying potential research questions based on literature reviews and available datasets.
2. Overview of research methodologies common in AI/ML, including experimental design and algorithm evaluation.

Offline Help: One-on-one proposal feedback sessions.

Optional Coding Bootcamp: Wrap-up session with Q&A addressing common coding challenges.

Weeks 5-8

Phase 2.

**Proposal Development
& Data Familiarization**

Week 5: Research Proposal Writing

1. Components of a strong research proposal (research question, objectives, methods, expected outcomes) with an AI/ML focus.

Offline Help: Personalized review and feedback on proposal drafts.

Week 6: Data Tools & Techniques

1. Introduction to data cleaning, visualization, and preliminary analysis using Excel or Python (libraries like Pandas and Matplotlib).

Offline Help: Hands-on practice with real-world datasets.

Week 7: Exploratory Data Analysis (EDA)

1. Conducting descriptive statistics and basic visualizations to uncover data trends and insights.

Offline Help: Guidance on interpreting analysis outcomes and refining research questions.

Week 8: Refining the Proposal

1. Incorporate insights from EDA into the research proposal.

Offline Help (60 min): Final review and sign-off on the research proposal.

Weeks 9-16

Phase 3.

**Data Analysis &
Research Report
Drafting**

Week 9-12: Data Analysis Workshops

1. Applying deeper EDA and statistical tests (e.g., correlation analysis, regression, hypothesis testing) and basic machine learning techniques where applicable.

Offline Help: Targeted support addressing specific data challenges and refining analyses.

Week 13-16: Advanced Data Analysis

1. Refinement of visualizations and ensuring robust statistical or model-based analysis.

Offline Help : Detailed one-on-one review of data interpretation and troubleshooting.

Weeks 17-22

Phase 4.

**Writing, Revision &
Final Presentations**

Week 17-19: Writing a Research Report

1. Focus on finalizing the written report, ensuring clarity, logical flow, and correct citation practices.

Offline Help: Intensive one-on-one editing, review of data presentation, and discussion on improving the report.

Week 20-22: Final Revision & Presentations

1. Finalize the research paper and develop a presentation (poster or slide deck) tailored for AI/ML projects.

Offline Help: Focused final edits and coaching on presentation skills.

Phase 1 Deep Dive

Foundations, Research Exploration, & Coding Basics

What is AI?

AI (Artificial Intelligence) is a branch of computer science that enables machines to perform tasks that typically require human intelligence

Learning: Machines can learn from data without explicit programming. Reasoning: AI systems can make decisions based on logic and probability. Problem Solving: AI can process data to find solutions.

Perception: AI can recognize images, understand speech, and more.

What is AI?

Learning: Machines can learn from data without explicit programming.

Reasoning: AI systems can make decisions based on logic and probability.

Problem Solving: AI can process data to find solutions.

Perception: AI can recognize images, understand speech, and more.

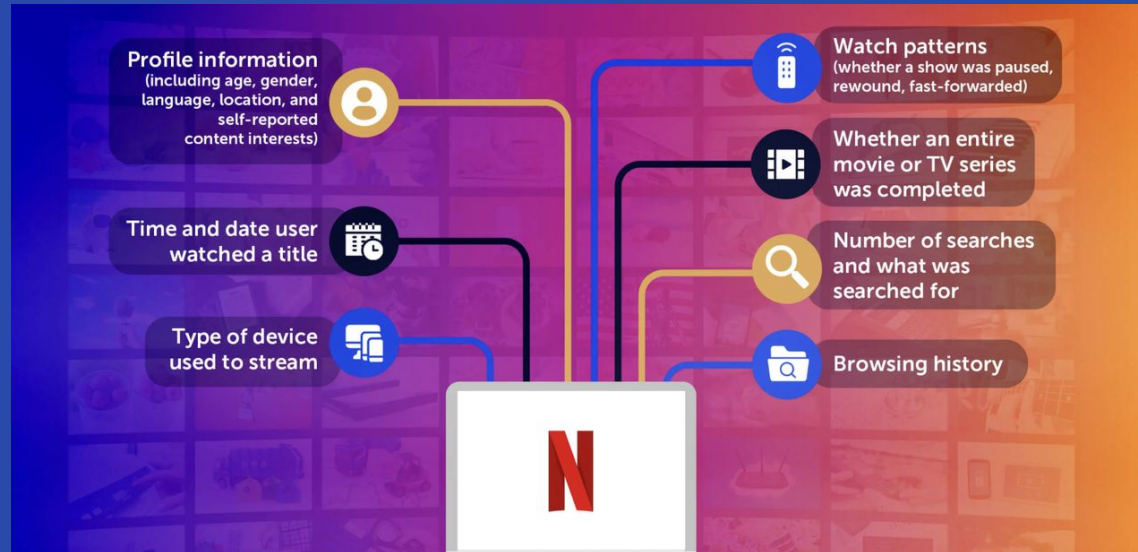
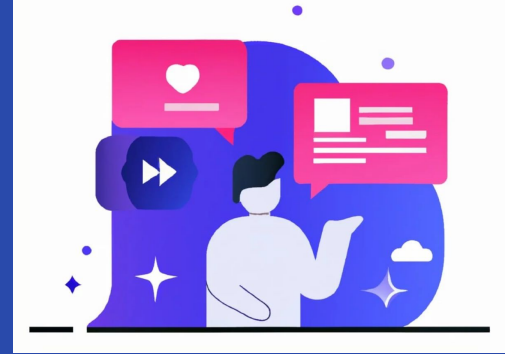
What is ML?

ML (Machine Learning) is a branch of Artificial Intelligence (AI) where computers learn from data to make predictions or decisions without explicit programming.

How It Works: ML systems identify patterns in data and improve their accuracy over time.

Social Media Feeds:

Personalized content recommendations



Virtual Assistants:



**Siri, Alexa,
Google Assistant**

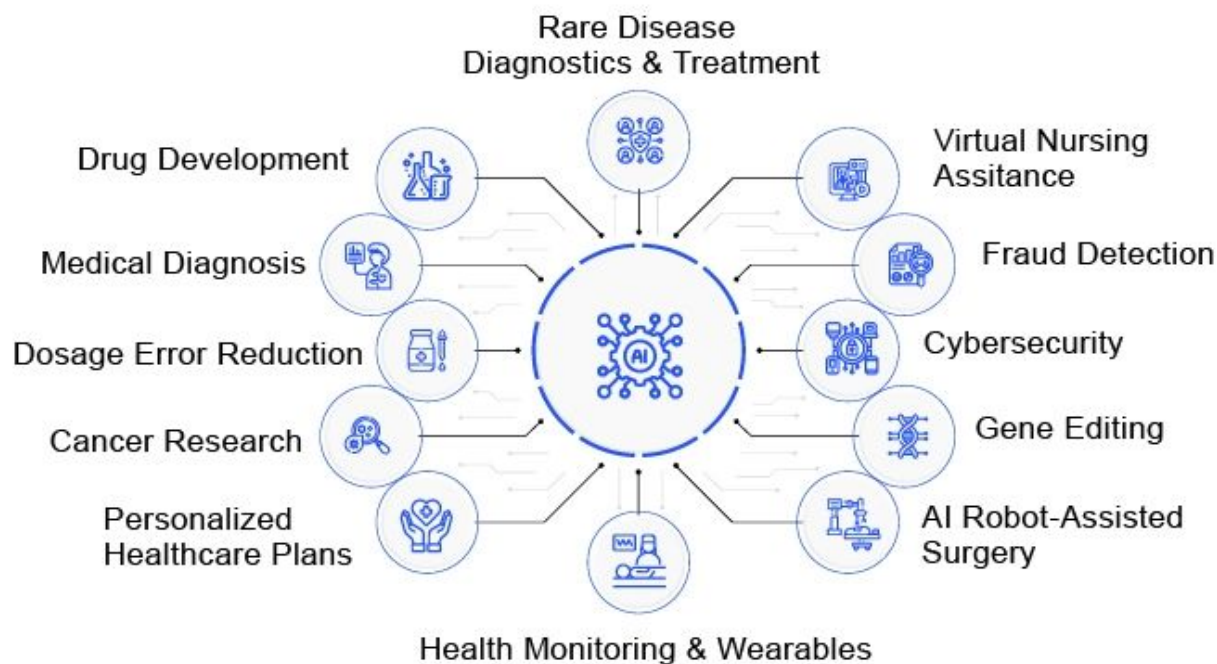


Chatbots: Customer service bots



Healthcare

Applications of AI in Healthcare



Fraud



Anomaly detection



Risk scoring



Network analysis



Text analysis



Identity verification



Adaptive learning

FaceID



Self Driving Cars



AI/ML and Research

Open Question: Can you think of other examples in your daily life where Artificial Intelligence (AI) or Machine Learning (ML) might be used?

Statistics and Research

Statistics allow us to see **trends**. Trends allow us to make **conclusions** that can be applied to **society**!



Example:

What is the efficacy of X drug?

Is X drug safe for usage?

Can the effects of X drug on one individual be assumed for other individuals?

Overview of the Research Process in AI/ML

- Literature Review: Research builds on previous work by analyzing existing studies, identifying gaps, and leveraging insights from prior methodologies. Tools like Google Scholar, arXiv, and AI-powered platforms can assist in summarizing and organizing relevant papers.
- Formulating a Hypothesis and Identifying a Research Question: Define the problem statement and research objectives clearly. A hypothesis provides a testable prediction based on observations or theories.

Overview of the Research Process in AI/ML

- **Data Gathering:** Sources for datasets include open repositories such as Kaggle, UCI Machine Learning Repository, Google Dataset Search, and domain-specific databases.
- **Analysis and Communication of Results:** Apply appropriate algorithms to analyze the data (e.g., **Random Forest**, **Logistic Regression**). Interpret results critically, validate findings, and communicate them through publications or presentations.

Step	Description
Literature Review	Analyze existing studies to identify gaps and build foundational knowledge
Hypothesis	Formulate a testable prediction and define research objectives
Data Gathering	Collect datasets from sources like Kaggle, UCI Repository, or synthetic data
Preprocessing	Clean and prepare data for analysis
Exploratory Analysis	Identify patterns and insights using statistical methods
Model Training	Train algorithms (e.g., SVM, Decision Trees) on prepared data
Validation	Evaluate model performance using metrics like accuracy or F1 score
Communication	Present findings through reports, papers, or visualizations

Methods of Research: Experimentation

Experimentation involves conducting controlled tests that manipulate **variables** to identify **cause/effect relationships**.

(ex: If I increase temperature, what happens to the stability of X protein?)

Case Studies

Case studies involve engaging in a deep dive into specific cases (stories) to observe unique phenomena.
(ex: Observing someone's experience with battling cancer.)

Case Study Example

https://colab.research.google.com/#scrollTo=hs7YnZ_pykHj&fileId=https%3A//storage.googleapis.com/kaggle-colab-exported-notebooks/scratchpad/energy-consumption-prediction-with-python.ccc546b2-69d0-43d1-a66e-7c07ca388c21.ipynb%3FX-Goog-Algorithm%3DGOOG4-RSA-SHA256%26X-Goog-Credential%3Dgcp-kaggle-com%2540kaggle-161607.iam.gserviceaccount.com/20250413/auto/storage/goog4_request%26X-Goog-Date%3D20250413T013614Z%26X-Goog-Expires%3D259200%26X-Goog-SignedHeaders%3Dhost%26X-Goog-Signature%3D314b226101bfb4ace81d3e10653bcc8d9e9fcda581c5a860dafa802ca1027c24bd7739aeb6654ad4c6f6d3988bb4cd93b5cb868f48d32a27c7d7072f81a4ca3643ad1497c2b3f3a6d9114e58f58fff6688d9c4e4e1581812d84c2be8cec467c0ceaf3250b915c5321fd1920d6cc3bb543a56d0ab708903eace6aed2deac22d02f85040877453e6591ab113c71e3fbb22406cc60c34d1284fa6645b372581fe5df366a6933db34fe9b677b4dc299dcea91e2fa5d6f776c9baf27d72d154338245aaaba6f3c06ca311aef63fd78d8117de99bc22a39d918e25c46a7dcc39e0e6dd76b11cef9d121999cade586210bf373efde2c57058feebf1ea940d7de9518d5ec

Statistics and Research

Open Question:

What step you find most interesting?

Methods of Research: Longitudinal Studies

Longitudinal studies involve data collection over a long span of time to observe long-term trends.

(ex: Tracking a group of people's growth as they take HGH (human growth hormone) over 15 years.)

Methods of Research: Surveys

Surveys involve taking a widespread census to collect general information from a large body of people.

(ex: Asking the state of CA to report their blood pressure levels.)

Breakout Brainstorm Activity

Question?

**Brainstorm potential research questions or projects
related to AI/ML?**

Recap

Program's vision:

1. Enable students to gain **AI/ ML research experience**
2. Different project pathways (conference paper, application, web project) based on interests.
3. Help students create works that look good on **college applications, resumes, etc.**

Basics of AI/ML

Artificial Intelligence (AI): The science of making machines "smart" by enabling them to simulate human intelligence.(Self-driving cars, Healthcare)

Machine Learning (ML): A subset of AI that teaches machines to learn from data and improve their performance over time without being explicitly programmed (Social media, Image Classifier)

Research Process: Literature Review → Hypothesis → Data Collection → Model Training → Evaluation

Literature Review: Analyze previous studies to understand existing knowledge and identify gaps.

Formulating Hypothesis: Define a clear research question or problem to solve.

Data Collection: Gather reliable data from sources like Kaggle, UCI Repository, or sensors. Ensure data is clean and relevant to the research question.

Model Training: Use algorithms to train models on the collected data (e.g., Linear Regression, Neural Networks).

Evaluation and Reporting: Validate model performance using metrics like accuracy or RMSE. Communicate results through visualizations, reports, or presentations.

Homework

Your Homework

[11:59 PM, 4/19 (Sat.)] Attend an offline meeting for help in
Discuss personal interests in AI and Q&A.

- Offline meetings can be scheduled on WhatsApp. A **Schej** (scheduling website) should be provided in a DM with you/your parent!

Feedback

Your Homework

[11:59 PM, 4/19 (Sat.)] In a Google Doc, list:

- a. 3-5 Research question in AI you'd be okay with writing about, along with an explanation for why you find it interesting (for each).**
- b. 1-3 questions you have about the AI research topics you listed.**
 - Make sure to provide access to the Google Doc!
 - Send to Ronel's ronelsolomon@gmail.com.

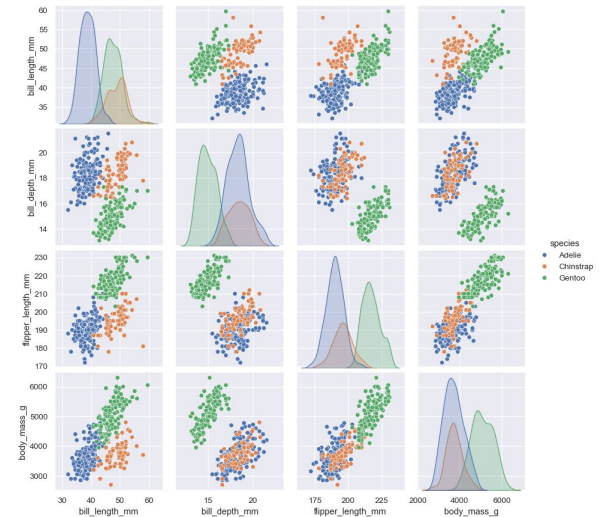
Key Takeaways

What will you learn from this program?

Student Takeaway: EDA

EDA (Exploratory Data Analysis) allows you to:

1. Visualize data **graphically**
2. Identify **summary statistics**
3. Identify **outlier data**
4. Check **data quality**



Student Takeaway: Computer Science

Computer science allows you to:

1. Work with programs that require **Python**
2. Branch off to **other coding languages**
3. Become familiar with **mathematical logic**
4. Explore future developments (**AI, modeling**, etc.)

Student Takeaway: Scientific Communication

Scientific communication allows you to:

1. Conduct online **research**
2. Identify a **thesis**
3. Develop logical **argumentation**
4. Present **findings** to a **wide audience**