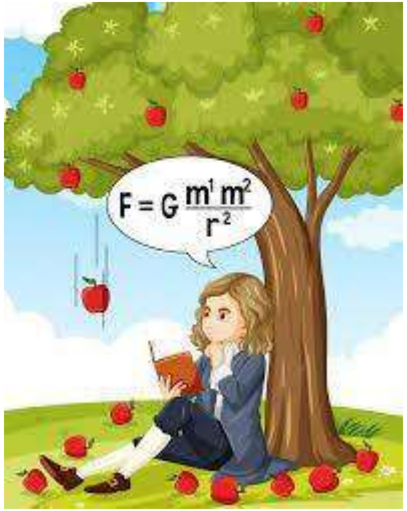
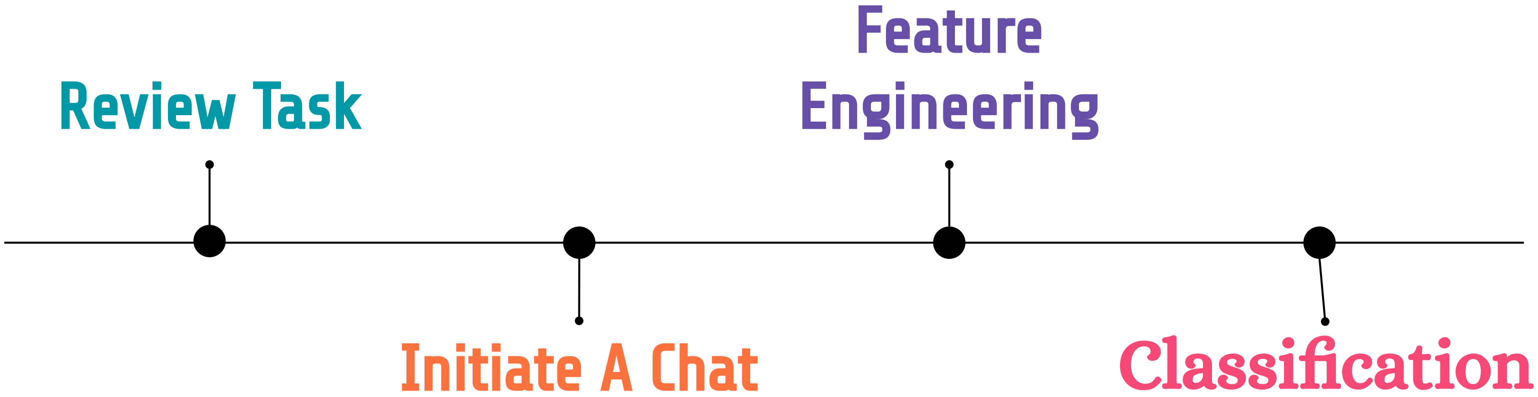
# Common Public Knowledge for Enhancing Machine Learning Data Sets

Arnon Ilani
Joint work with Shlomi Dolev
Ben Gurion University

Machine learning ignores thousand of years knowledge…



# Research objectives and flowchart



Review Task

Initiate A Chat

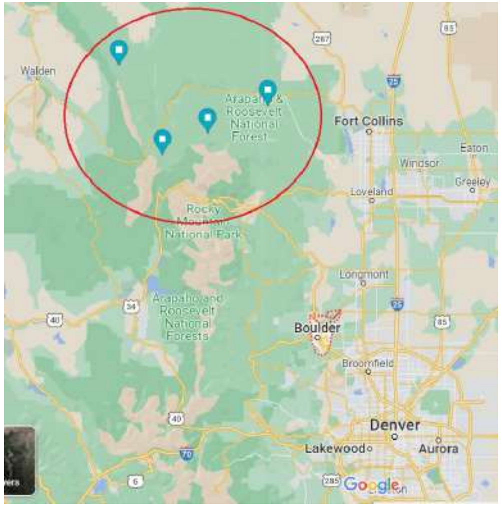
Feature Engineering

Classification

# THE COVERTYPE DATASET

## Task Review:

1. Classes: 7 tree types.
2. Features: 12 including:Elevation, USFS ELU Code soil type, etc.
3. Domains: Geology, Geography, Chemistry, Biology, Botany, Topography, Soil Science, Exposure to sunlight, Man-made parameters: nearby roads, wildfire ignition points.


Walden
Arapaho & Roosevelt National Forest
Fort Collins
Eaton
Windsor
Greeley
Loveland
Rocky Mountain National Park
Longmont
Arapaho and Roosevelt National Forests
Boulder
Broomfield
Denver
Lakewood
Aurora
Google

## Task Goal:

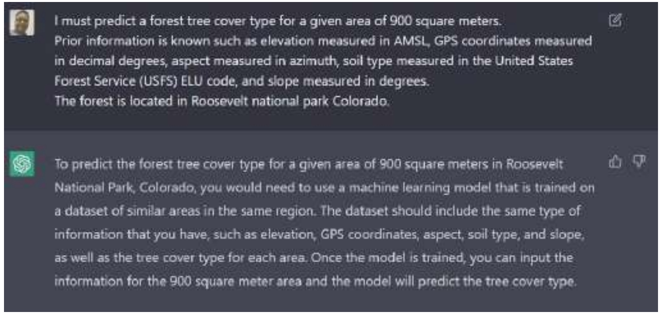
Tree Type Classification, Size: 581,012

Covertype by Jock Blackard 1998.

# THE COVERTYPE DATASET

## Initiate A Chat:

1. **Describe task to frame the ChatGPT with as much details.**
2. **The straightforward approach does not work usually.**
3. **Starting with guiding questions until we get recommendations:**
4. **Use soil science.**
5. **Use Soilweb app by CSRL.**



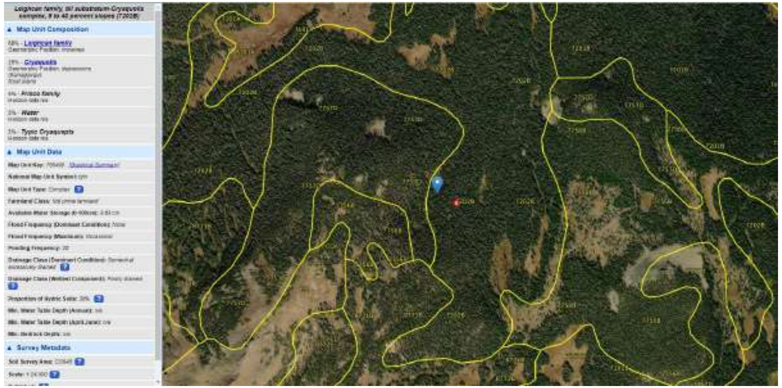
I must predict a forest tree cover type for a given area of 900 square meters.
Prior information is known such as elevation measured in AMSL, GPS coordinates measured in decimal degrees, aspect measured in azimuth, soil type measured in the United States Forest Service (USFS) ELU code, and slope measured in degrees.
The forest is located in Roosevelt national park Colorado.

To predict the forest tree cover type for a given area of 900 square meters in Roosevelt National Park, Colorado, you would need to use a machine learning model that is trained on a dataset of similar areas in the same region. The dataset should include the same type of information that you have, such as elevation, GPS coordinates, aspect, soil type, and slope, as well as the tree cover type for each area. Once the model is trained, you can input the information for the 900 square meter area and the model will predict the tree cover type.

# THE COVERTYPE DATASET

## Initiate A Chat:

1. **Zoom in on:**
2. **New Feature: Soil drainage.**
3. **New Feature: Soil erosion.**
4. **Reaching a limit.**



Leighcan family, till substratum-Cryaquolls complex, 0 to 40 percent slopes (72028)

▲ Map Unit Composition

68% - Leighcan family
Geomorphic Position: [illegible]

19% - Cryaquolls
Geomorphic Position: depressions
[illegible]

6% - Frisco family
[illegible]

2% - Water
[illegible]

3% - Typic Cryaquepts
[illegible]

▲ Map Unit Data

Map Unit Key: [illegible] [illegible]

National Map Unit Symbol: [illegible]

Map Unit Type: Complex

Farmland Class: Not prime farmland

Available Water Storage (0-100cm): [illegible]

Flood Frequency (Dominant Condition): None

Flood Frequency (Maximum): Occasional

Ponding Frequency: 20

Drainage Class (Dominant Condition): Somewhat excessively drained

Drainage Class (Wettest Components): Poorly drained

Proportion of Hydric Soils: 20%

Min. Water Table Depth (Annual): n/a

Min. Water Table Depth (April-June): n/a

Min. Bedrock Depth: n/a

▲ Survey Metadata

Soil Survey Area: [illegible]

Scale: 1:24,000

https://casoilresource.lawr.ucdavis.edu/gmap/

# THE COVERTYPE DATASET

## Feature Engineering:

1. Original features: Elevation.
2. New features: 13 in total. E.g., Available Water Storage, Total Plant Available Water, T-Erosion Factor, etc.

## Classification Results:

We found that the accuracy level improved by 1-4% from 67% to 68-71%.

# THE ISOLATED LETTER DATASET

## Task Review:

1. Classes: 26 English letters
2. Features: 616 audio features, Age, Gender, US State.
3. Domains: custom engineered features, Age, Gender, US State, Phonetics, Sociology, Sociophonetics

## Task Goal:

English letters classification, Size: 7,797

The Isolated Letter (ISOLET) Speech Recognition dataset 1994.

# THE ISOLATED LETTER DATASET

## Feature Engineering:

1. **Original features: two audio features.**
2. **Chat recommendation: Labov's DSUS**
3. **New features: 5 in total. E.g., /aw/ fronting, Glide deletion of /ay/, etc.**

F2 of /aw/
● > 1200
● > 1450
● > 1650
● > 1750
West
Canada
North
Midland
South

Map 20.1. Fronting of /aw/ in the West

## Classification Results:

**An improvement of up to 3% in the average classification accuracy across the 26 classes.**

# Conclusions

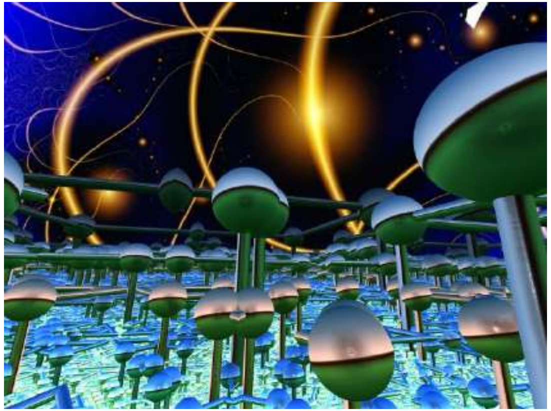
## Good but not great

1. For the Forest Cover Type dataset, we demonstrated an improvement in average classification accuracy.
2. We found limited usefulness of soil type for predicting tree types.
3. For the ISOLET Speech Recognition dataset, leveraging sociolinguistics and sociophonetics research was more complex than anticipated.
4. Using a large language model like ChatGPT as an expert knowledge source showed benefits.

# Concerns and future work

## The future will be great

1. **Making manual decisions and feature engineering.**
2. **The propagation of misinformation in ChatGPT.**
3. **The focus is on demonstrating the presence of readily accessible knowledge rather than providing explicit dataset recommendations.**



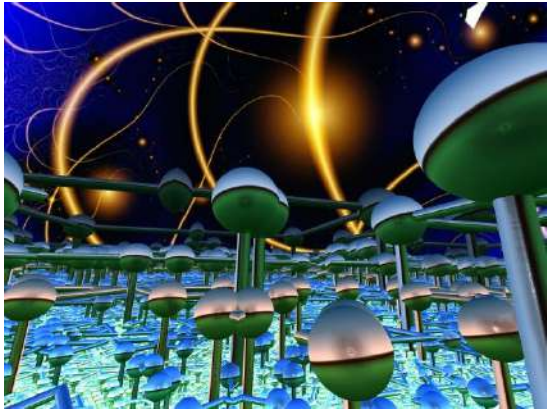
## Five months later:

1. **GPT-4 Technical Report, SOTA in WinoGrande and MMLU, 27 Mar, 2023**
2. **Large Language Model Guided Tree-of-Thought, Jieyi Long, 15 May, 2023**
3. **PaLM 2 Technical Report, Google, 17 May, 2023**

# Concerns and future work

## The future will be great

**By utilizing ChatGPT as the domain expert, an expert system can be constructed, leveraging its machine learning capabilities to attain improved outcomes.**



# Thanks!

**Questions?**

**arnonil@post.bgu.ac.il**