Navigating the Integration of Machine Learning into Domain Research

Bernie Boscoe
Assistant Professor of Computer Science
Southern Oregon University

Was an RSE before the name

7 years ago I started working with astronomers on machine learning projects

Starting to work with environmental scientists on a land bridge project

I research infrastructures for machine learning for scientific research

My current goal is to develop undergraduate training for RSE roles

Goal of this talk: “Food for thought” on getting started with ML/AI projects for domain research from an RSE perspective

This work started as a Sloan Foundation funded project to understand what it means for astronomers to adopt machine learning practices, evolving into domain sciences adopting ML (our team is at UCLA)

Methods: scores of interviews with domain researchers, RSEs, ethnography, I’m going to present the common themes in an informal way

Intended audience: less resources on hand, RSEs must wear many hats, or perhaps you have ⅛ of an RSE in your group

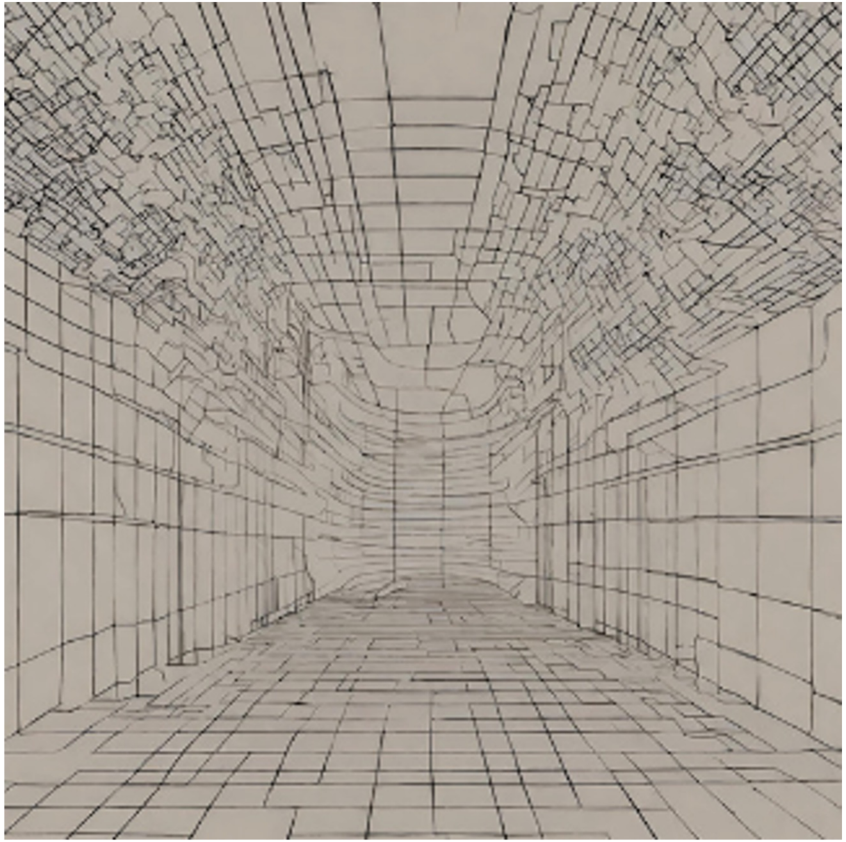
For fun I thought I’d use text to image AI generation to illustrate my slides

Each image on each slide uses the text above as its prompt

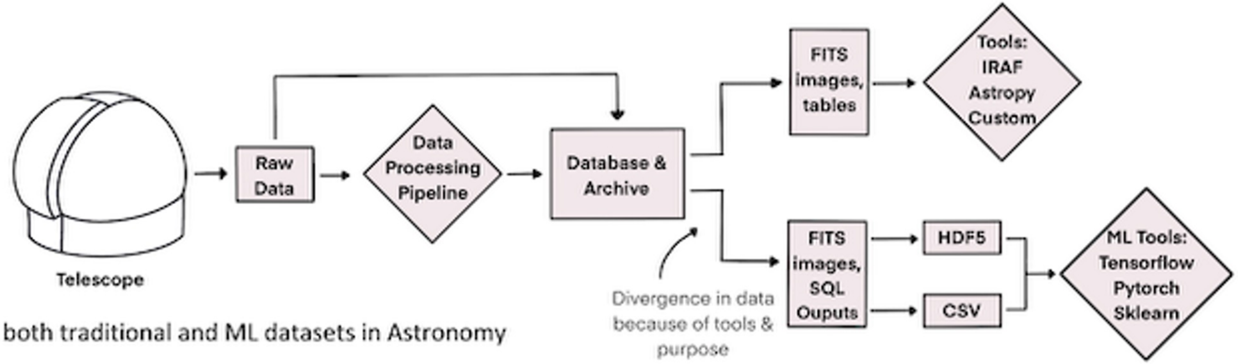
Prompt: **a Research Software Engineer wearing many hats** (!)

Prompt: **What is the data? Size, file formats, is it 'machine learning ready' (for ingestion for a model or task)**

Prompt: How are the **versions of data** and models going to be kept track of?



Prompt: How are the **versions of data** and models going to be kept track of?



Divergence in data because of tools & purpose

Data flow for both traditional and ML datasets in Astronomy

Prompt: Where are you going to train your model? (Commercial cloud platform, NSF-sponsored cloud, build your own, combination)

– Commercial cloud pricing is nebulous, and for students can be very problematic

–NSF platforms good for pricing but learning curve (‘nix, scripting, slurm, MPI)

–Roll your own Threadripper (2 **NVIDIA GeForce RTX 4090s)**

Prompt: What are benchmarks domain scientists will use to determine an acceptable model?
(Scientific ground truth domain norms + common ML “accuracy” reporting)

Prompt: Speed of ML/AI in CS far faster than in domain sciences– to do good science we need to document code, experiments, processes– ‘slow ML movement’-- we need stability for science

Prompt: Thank you US-RSE and I’m so excited for the first US-RSE Conference!!

boscoeb@sou.edu