MLM25 Sprint 1 +
Exploratory Data Analysis & Reproducible ML

Slides: go.wisc.edu/7w28y9

WISCONSIN INSTITUTE FOR DISCOVERY

data science HUB

ML+X

Slides: go.wisc.edu/7w28y9

Tonight's Agenda

1. Sprint (teamwork time)
2. Know Your Data (EDA)
3. Know Your Experiments (Reproducible ML)
4. Sprint (teamwork time)

Slides: go.wisc.edu/7w28y9

Sprint (Teamwork til 5pm)

1. **Team registration due by midnight TONIGHT**
   a. https://forms.gle/UrmRK7q3CZjuqUZz6 – **Only ONE person** from each team should fill out the form

2. **Team report outs – please add responses here go.wisc.edu/82y993**
   a. **We'll discuss the prompts more in detail at 5pm.**

Before throwing 100 diff. models at the problem...

- **What should you do first?**

KNOW YOUR DATA !

Garbage in = Garbage out

Our models are only as "intelligent" as our training data.

Good ML practice starts with **understanding your data** through exploratory data analysis (EDA).

- **Reveal structure**: Distributions, correlations, clusters, etc.
- **Expose issues**: Missing values, bias, noise, mislabels, outliers.
- **Relate to modeling**: Which features/signals look promising? Where might noise be a problem?
- ***Iterative → revisit EDA as models reveal new gaps.***

What's one idea your team has for initially exploring your data?

Add your team's thoughts:
go.wisc.edu/82y993

Slides: go.wisc.edu/7w28y9

Example EDA Notebook: Titanic Dataset

Need additional guidance and inspiration on how to start your EDA?
Follow along with this Google Colab notebook to see some examples working with the Titanic dataset

EDA notebooks are encouraged as Nexus posts! Notebook should include:

- Short justification of each step
- Insights gained at each step
- Commented code

Email endemann@wisc.edu with Colab notebooks for Nexus :)

Before throwing 100 diff. models at the problem...

- **What should you do first?**

ESTABLISH BASELINE !

Baseline Models

- Fast to build, easy to understand
- Often reveals issues in the data before you sink time into big models.
- Provides a reference point to improve from

Examples

- If predicting future rainfall, use today's rainfall as prediction.
- Linear regression BEFORE a 100B parameter LLM
- If you need LLMs (e.g., RAG), use a smaller one as baseline

What's the simplest baseline you can run this week (or next) that helps you understand your data and gives you something to improve on?

Add your team's thoughts:
go.wisc.edu/82y993

Slides: go.wisc.edu/7w28y9

Before throwing 100 diff. models at the problem...

- **What should you do first?**

PRACTICE REPRODUCIBILITY !

Slides: go.wisc.edu/7w28y9

Tips for Reproducible ML

- **Version control** → use GitHub; commit often; Kaggle commits auto-save.
  - Avoid tracking notebooks directly (large commit diffs)
- **Environment tracking** → requirements.txt, environment.yml, lockfiles.
  - Kaggle notebooks come pre-loaded with ML libraries
  - For local setups, try `uv` as a package manager – video tutorial.
- **Data discipline** → keep raw/ vs processed/ separate.
- **Experiment logs** → record dataset version, params, metrics.
  - manual logging, MLFlow, or Weights and Biases
- **Collaboration habits** → document decisions (e.g., in README), share notebooks often
- **Extras** → set random seeds, save models with clear names.

How will your team make your work reproducible so that others — including your future self — can build on it?

Add your team's thoughts:
go.wisc.edu/82y993

Slides: go.wisc.edu/7w28y9

MLM25 — Looking Ahead

1. **9/25 (Thur), 4:30-6:30pm**: Sprint 2 + U-Net Demo

2. **10/2 (Thur), 5:30-7:30pm**: **Exploratory data analysis presentations**

Full schedule: ml-marathon.wisc.edu/schedule/

Slides: go.wisc.edu/7w28y9

EDA Presentations on 10/2, 5:30-7:30pm

- With ~20 teams, each team will have 4 minutes to present
  1. **Introduce your data**: Briefly explain what kind of data you're working with.
  2. **Highlight key steps**: Mention the most important data exploration or cleaning actions you've taken.
  3. **Useful tools/packages/functions**: Mention any useful tools/libraries you used for your analysis
  4. **Share insights**: Discuss any early patterns or challenges you've discovered so far.
  5. **Baseline model**: Discuss results of baseline model. How hard is the task?
  6. **Next steps**: Include ideas for next steps

- Send google slides link (5 slides max) by **9/30, 11:59pm** to endemann@wisc.edu.
  - **Format slides as Widescreen 16:9** (file -> page setup)
  - **Synced slides**: Slides can be *polished* up until presentation on 10/2. However...
    - No rearranging, adding, or removing slides after 9/30. These changes will not sync!

Slides: go.wisc.edu/7w28y9

Sprint (Teamwork til 6:30pm)

1. **Team registration due by midnight TONIGHT**
   a. https://forms.gle/UrmRK7q3CZjuqUZz6 – **Only ONE person** from each team should fill out the form

2. **Team report outs – please complete responses by 6:30pm go.wisc.edu/82y993**