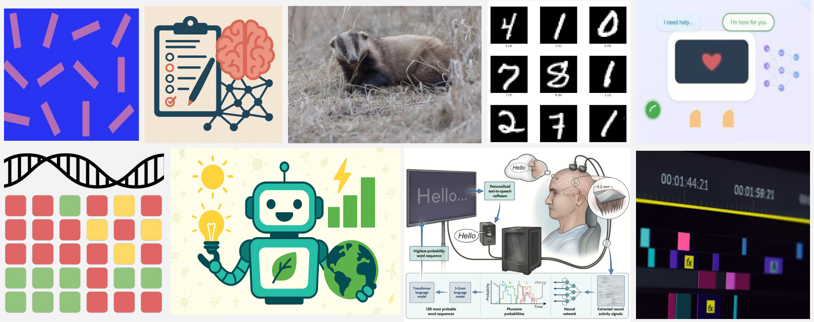
Welcome To The Machine Learning Marathon 2025 (MLM25)!

Slides: **go.wisc.edu/2et35u**

WISCONSIN INSTITUTE FOR DISCOVERY

data science HUB



ML+X

Slides: go.wisc.edu/2et35u

Thanks, Organizing Team!

- **Abrar Majeedi – Graduate Student, BMI | UW-Madison**
  Presenter
- **Adam Ross Nelson – Teaching Faculty, Psych | UW-Madison**
  Project Organizer, Advisor
- **Alan McMillan – Professor, Radiology | UW-Madison**
  Leadership, Advisor
- **Annie Zhao – Undergraduate Student | UW-Madison**
  Project Organizer
- **Carl Kashuk – Data Scientist, Radiology | UW-Madison**
  Presenter
- **Chris Endemann – Research Cloud Consultant | RCI**
  Leadership, Project Organizer, Advisor, Presenter
- **Dhruba Jyoti Paul – ML Engineer | ALL3D**
  Project Organizer, Presenter
- **Eli Cytrynbaum – Graduate Student | UW-Madison**
  Project Organizer
- **Nick Kyburz – AI/MLOps Technical Specialist | IBM**
  Presenter
- **Ross Jacobucci – Professor | Center for Healthy Minds**
  Project Organizer
- **Yin Li – Professor, BMI | UW-Madison**
  Leadership, Advisor
- **Yuriy Sverchkov – Scientist, BMI | UW-Madison**
  Project Organizer, Advisor

Thank You, ML+X Sponsors!

WARF
Wisconsin Alumni Research Foundation

aws

Promega

Thanks GCP for Extra Compute Power!

Google Cloud

Slides: go.wisc.edu/2et35u

Tonight’s Agenda

1. Motivation
2. Logistics
3. Overview of projects
4. Find your team!
5. Resources for getting started

Slides: go.wisc.edu/2et35u

Motivation – Why MLM25

- **Rapid exposure to diverse methods and application areas**. Sharing across projects sparks insights no single project could.
- **A testbed for real use cases**: Challenges mirror ML/AI problems at UW–Madison, industry, and beyond
- **ML/AI is empirical:** Real skill comes from applying methods to diverse datasets and problems, not from theory alone.
- **Career-building**: A portfolio of projects opens doors of this field.
- **Community-driven:** This isn't a class — it's collaborators building together. Lean on each other for help!

Shared failures & successes create skills and connections that outlast the hackathon.

Slides: go.wisc.edu/2et35u

MLM25 Code Of Conduct

1. **Respect and Inclusivity**: Encourage a diverse range of perspectives and backgrounds. Avoid offensive language or comments that might make others uncomfortable.

2. **Be Patient and Supportive**: Be patient with individuals who may not have as much experience. Offer assistance and mentorship when appropriate and requested.

3. **Embrace the Kaggle Spirit!** Participants should actively contribute to the community by sharing their work and insights (via Kaggle, Slack, meetings, etc.)

4. **Feedback and Critique**: Provide constructive feedback when necessary, but do so in a respectful and considerate manner. Be open to receiving feedback on your own work and ideas.

If you encounter any code of conduct violations, email me (endemann@wisc.edu).
I will work with you to figure the next best steps.

Rules

1. **Team Size**: The maximum team size is 5. Teams with only one member are highly discouraged, but some exceptions may be granted.

2. **One Challenge Per Participant**: To ensure efforts are not spread too thin.

3. **Challenge Specific Rules**: Be sure to review the "Rules" section of your chosen challenge to see if any additional rules may apply.

4. **Attendance**: Occasional absences (e.g., for exams or unavoidable conflicts) are understandable, but please <u>be respectful of your fellow teammates and the commitment they've made to fully participating in MLM25</u>.

5. **Challenge Specific Rules**: Be sure to review the "Rules" section of your chosen challenge (on Kaggle) to see if any additional rules may apply.

Slides: go.wisc.edu/2et35u

Digit Recognizer: Getting Started with Deep Learning

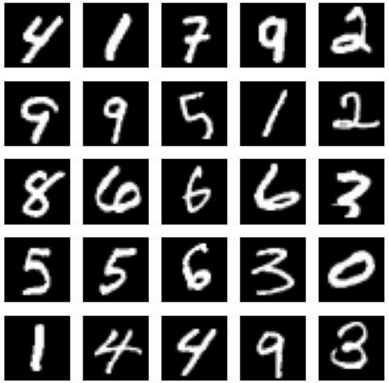
- **Goal**: Build and train neural networks to recognize digits from image data, exploring core concepts like model architecture, loss functions, and evaluation. *Teams are encouraged to implement solutions in multiple frameworks (e.g., PyTorch and Keras)* to deepen their understanding and compare tooling. While this challenge is more study group than full-blown hackathon, it's a strong foundation for more complex work in future semesters.

- **Prerequisites**: Some experience with R or Python and machine learning basics

- **Methods**: Deep learning and computer vision

- **Resources for Getting Started**
  - Also see **Intro to PyTorch** & **Intro to Keras**

- **Organizer/Contact**: NA (hosted by Kaggle). See Discussion tab from challenge page to get help.



Slides: go.wisc.edu/2et35u

WattBot: Estimating AI Emissions & Costs with RAG

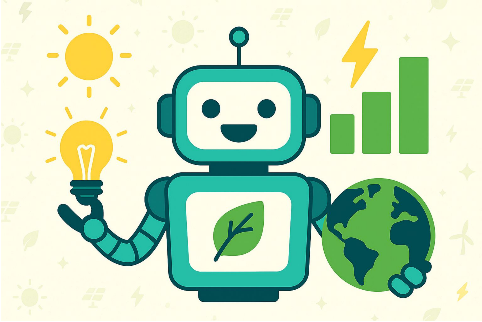
- **Goal**: In this challenge, teams will build a retrieval augmented generation (RAG) system that can reliably extract insights about AI's environmental impact across 30+ academic papers and industry reports. The system must output citation-backed answers—or explicitly indicate when the evidence is missing from the provided corpus

- **Prerequisites**: Prior exposure to NLP and deep learning is recommended. No prior experience with large language models (LLMs), RAG, or Hugging Face is required, but you must be willing to learn!

- **Methods:** RAG, LLMs, Text mining, Deep learning, Optical character recognition (optional)

- Resources for Getting Started

- **Organizer/Contact**: Chris Endemann (endemann@wisc.edu)

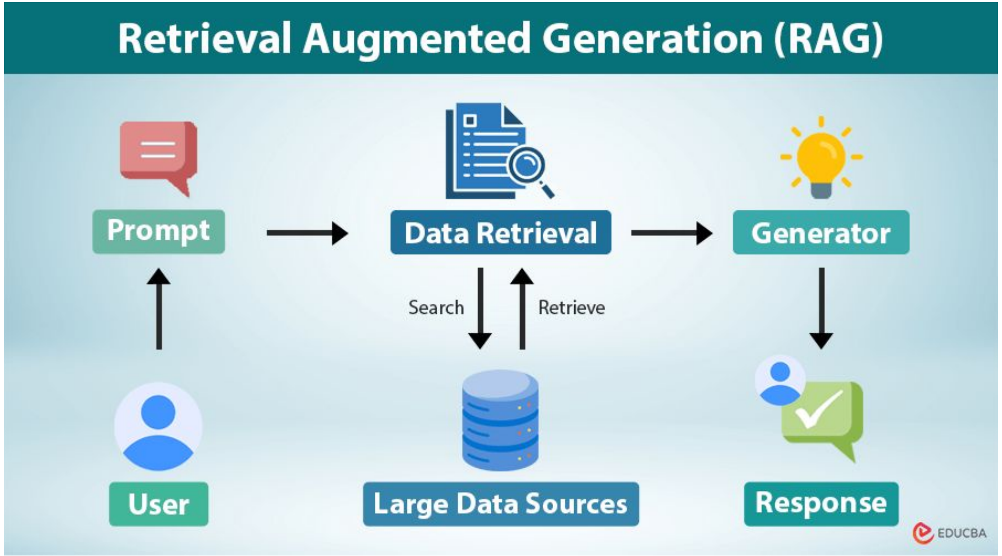


Slides: go.wisc.edu/2et35u

WattBot: Estimating AI Emissions & Costs with RAG

1. Given a question, use a large language model (LLM) to scan corpus/docs for relevant info (quotes, figures, tables, etc.)

2. Then, generate an answer using the retrieved info as context. This helps prevent "hallucinations" and makes responses grounded by citable materials.



https://www.educba.com/retrieval-augmented-generation/

Some questions are simple quote lookups

- ***True or False: New AI data centers often rely on air cooling due to high server power densities.***
  - **Answer**: FALSE.
  - **Supporting refID(s)**: [li2025b]
  - **Supporting material(s)**: [“In general, new data centers dedicated to AI training often rely on liquid cooling due to the high server power densities.”]

Some questions require reasoning or simple math.

- ***How many U.S. household-years of electricity consumption is training a 6.1B-parameter language model equivalent to?***
  - **Answer**: 1.3 household-years.
  - **Supporting refID(s)**: [dodge2022, strubell2019]
  - **Supporting materials**: ["The 6.1B parameter model consumed 13.8 MWh."; "Average U.S. household consumption ≈ 10,715 kWh/yr."]
  - ***Reasoning***: 13.8 MWh ÷ 10.7 MWh/yr ≈ 1.3 household-years of electricity.

Slides: go.wisc.edu/2et35u

Some questions require you to properly parse table data from PDFs

- *What is the estimated CO2 emissions (in pounds) from training the BERT-base model for 79 hours on 64 V100 GPUs?*

  Answer: 1438 lbs. **Supporting refID(s)**: [strubell2019], **Supporting materials**: Table 3.

Model	Hardware	Power (W)	Hours	kWh·PUE	CO_2e	Cloud compute cost
$\text{Transformer}_{base}$	P100x8	1415.78	12	27	26	\$41–\$140
Transformer_{big}	P100x8	1515.43	84	201	192	\$289–\$981
ELMo	P100x3	517.66	336	275	262	\$433–\$1472
BERT_{base}	V100x64	12,041.51	79	1507	1438	\$3751–\$12,571
BERT_{base}	TPUv2x16	—	96	—	—	\$2074–\$6912
NAS	P100x8	1515.43	274,120	656,347	626,155	\$942,973–\$3,201,722
NAS	TPUv2x1	—	32,623	—	—	\$44,055–\$146,848
GPT-2	TPUv3x32	—	168	—	—	\$12,902–\$43,008

Table 3: Estimated cost of training a model in terms of CO_2 emissions (lbs) and cloud compute cost (USD).[7] Power and carbon footprint are omitted for TPUs due to lack of public information on power draw for this hardware.

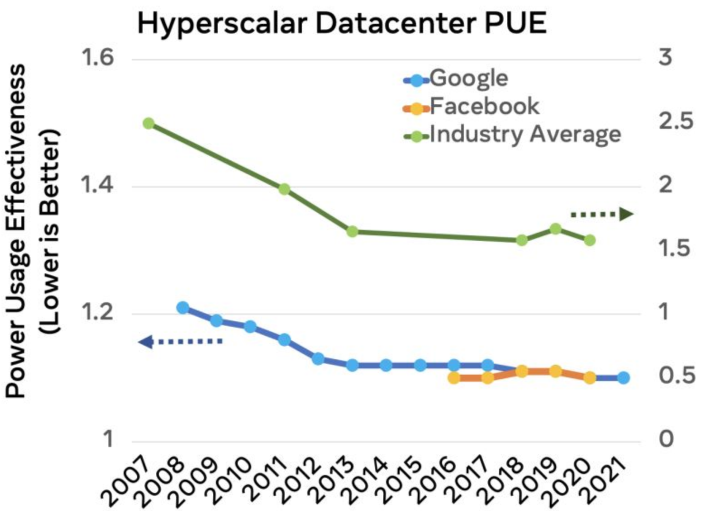
https://arxiv.org/abs/1906.02243

Slides: go.wisc.edu/2et35u

Some questions require visual reasoning over figures

- *True or False: In 2020, Power Usage Effectiveness (PUE) values reported by Google and Facebook (~1.1) were lower than the industry average by about 40%.*

  Answer: TRUE. **Supporting refID(s)**: [wu2021b], **Supporting materials**: [Figure 1]



https://arxiv.org/abs/2108.06738

Slides: go.wisc.edu/2et35u

Some questions are intentionally out-of-scope. System must respond appropriately.

What was the total electricity consumption of all Google Cloud TPU pods worldwide in 2023, in megawatt-hours?

- **Answer**: Unable to answer with confidence based on the provided documents.
- **Supporting refID(s)**: [is_blank]
- **Supporting materials**: [is_blank]

Slides: go.wisc.edu/2et35u

MaveDB amino acid substitution prediction

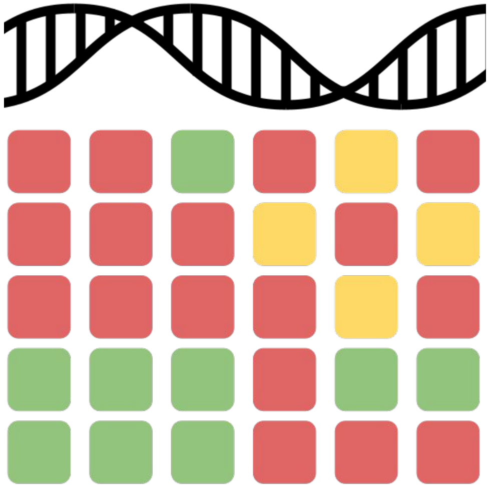
- **Goal**: Predicts the outcomes of Multiplexed Assays of Variant Effects (MAVEs). Given an amino acid substitution in a protein, you will predict the measurement of protein function that one would get if the protein were assayed in the lab.

- **Prerequisites:**
General ML understanding, familiarity with deep learning highly recommended, experience using Hugging Face is a plus.

- **Methods:**
Deep learning, Foundation models, Zero-shot prediction, Protein language models, Regression

- **Resources for Getting Started**
  - The huggingface model hub is a great resource for accessing pre-trained models
  - Zero-shot classification: Using large pre-trained models on novel tasks. These resources are written with a focus on LLMs, but many of the same ideas apply to protein language models.
  - ESM Cambrian: a state-of-the-art protein language model.

- **Organizer/Contact**: Yuriy Sverchkov (sverchkov@wisc.edu)



Slides: go.wisc.edu/2et35u

MaveDB amino acid substitution prediction

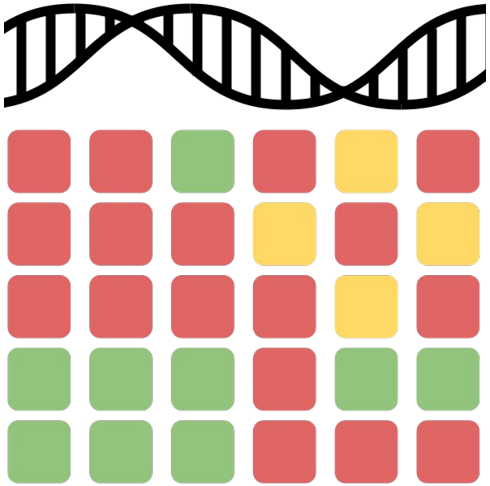
- **Methods:** A promising approach to this problem would be to use a trained “foundation” protein language model and use it as a feature generator for a trained ML “head.”

Proteins

- Proteins are the molecular machines that living cells rely on
- Made up of chains of amino acid residues
- Can be represented as a sequence of symbols, where each symbol represents an amino acid:

  MPLYSVTVKWGKEKFEGVELNTDEPPMVFKAQLFALTGVQP

- What happens to the protein’s function if one of the amino acid residues is substituted for another?



Protein Language Models

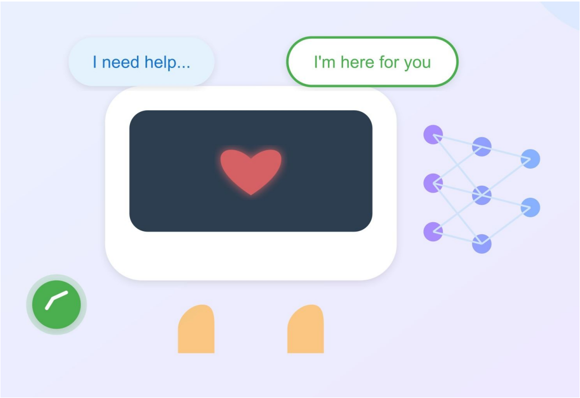
- Just as LLMs can model language as a sequence of tokens, PLMs can model proteins as sequences of amino acids
- Have been used for variant effect prediction and protein engineering

Slides: go.wisc.edu/2et35u

CrisisCompanion: Creating a Mental Health Support Chatbot

- **Goal**: In this challenge, you'll explore how generative AI can be responsibly applied to the development of mental health support chatbots. You'll work with a range of publicly available datasets—from counseling transcripts to emotion classification corpora—to build and evaluate a conversational agent that can provide basic, non-clinical support.

- **Long-Term Goal:** A deployable, real-time chatbot used in suicide prevention studies.
  - Oriented to suicide research (Needs to be able to discuss suicide)
  - Deployable in a restricted data environment
  - Understands its limits 6



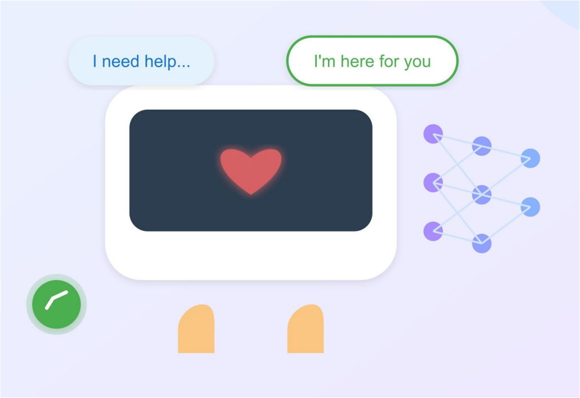
Slides: go.wisc.edu/2et35u

CrisisCompanion: Creating a Mental Health Support Chatbot

- **Prerequisites**: None – Use Claude/ChatGPT/Gemini to get started

- **Quick Start Recommendations:**
    - See Hugging Face's *transformers* quick start
    - Go through chatbot and building LLM tutorials
    - Qwen3 is a good starting place for an LLM

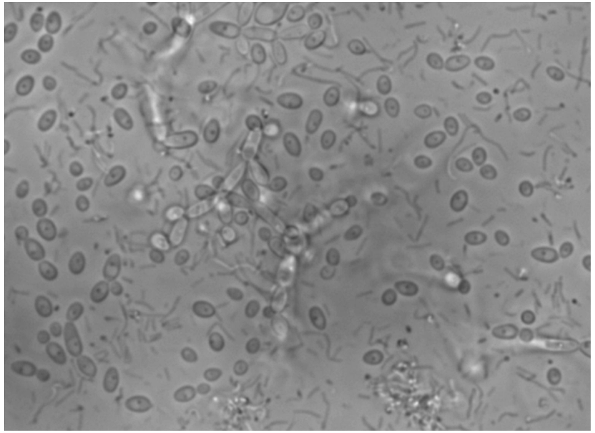
- **Organizer/Contact**: Ross Jacobucci (jacobucci@wisc.edu)

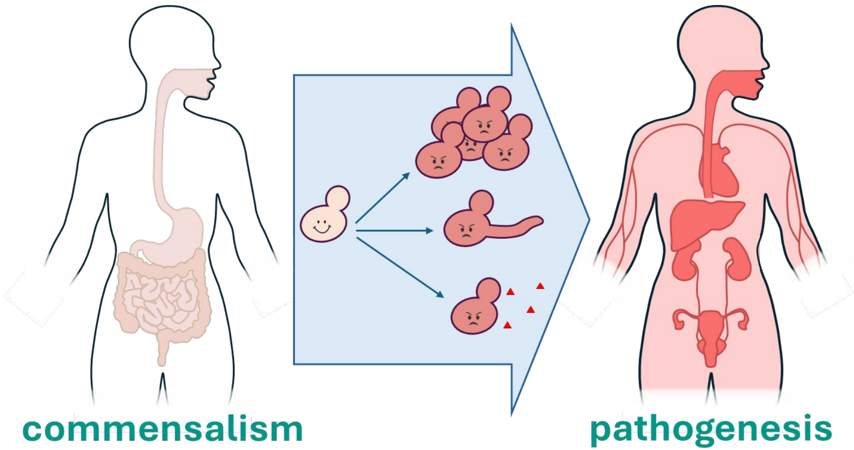


Slides: go.wisc.edu/2et35u

Cellular state image analysis

- **Goal**: Identify cells in different cellular states within time-lapse microscopy images.
- **Background**: *Candida albicans* is a fungal opportunistic pathogen in which both commensal and pathogenic behavior relies heavily on transitions between cellular states including yeast-form and hyphal cells and community organizations including dispersed cells and biofilms.





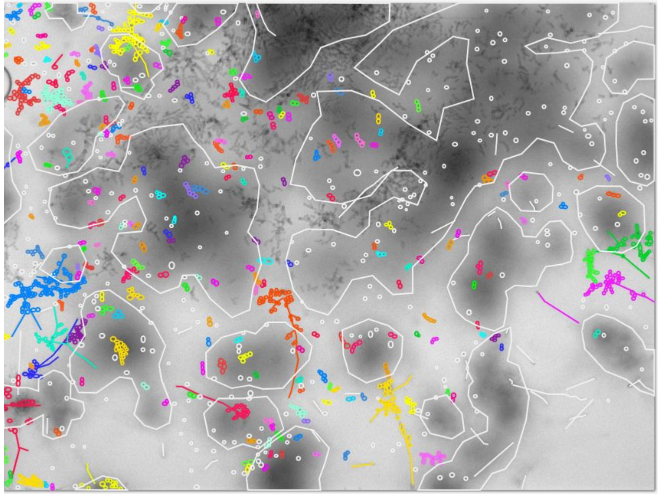
Slides: go.wisc.edu/2et35u

Cellular state image analysis

- Teams will be supplied with annotated and unannotated time-lapse images of developing and dispersing *Candida albicans* biofilms.

  Annotations include specification of:
  - cell type
  - cell dimensions
  - groups of adherent cells.

- Teams will use this data to create programs to:
  - identify cells
  - classify cells
  - track cells



- **Organizer/Contact**: Eli Cytrynbaum (cytrynbaum@wisc.edu)

Slides: go.wisc.edu/2et35u

Clustering the BioTrove Dataset

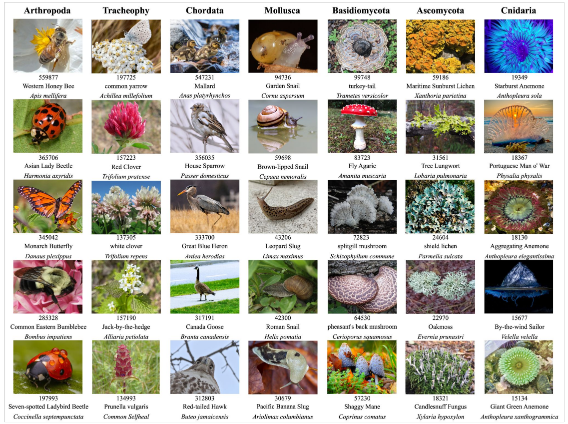
- **Goal**: In this challenge, you'll work with a ~50,000-image subset of BioTrove spanning the tree of life. Your task is to learn clusters from image features (with optional family-level taxonomic labels) that approximate the hidden biological taxonomy. **Evaluation is based on how well your clusters match hidden genus- and species-level groupings, measured using Normalized Mutual Information (NMI)**.

- **Prerequisites**: Familiarity with neural networks and CNNs is expected. You don't need to know contrastive learning or autoencoders beforehand, but these methods are encouraged and can be learned during the challenge.

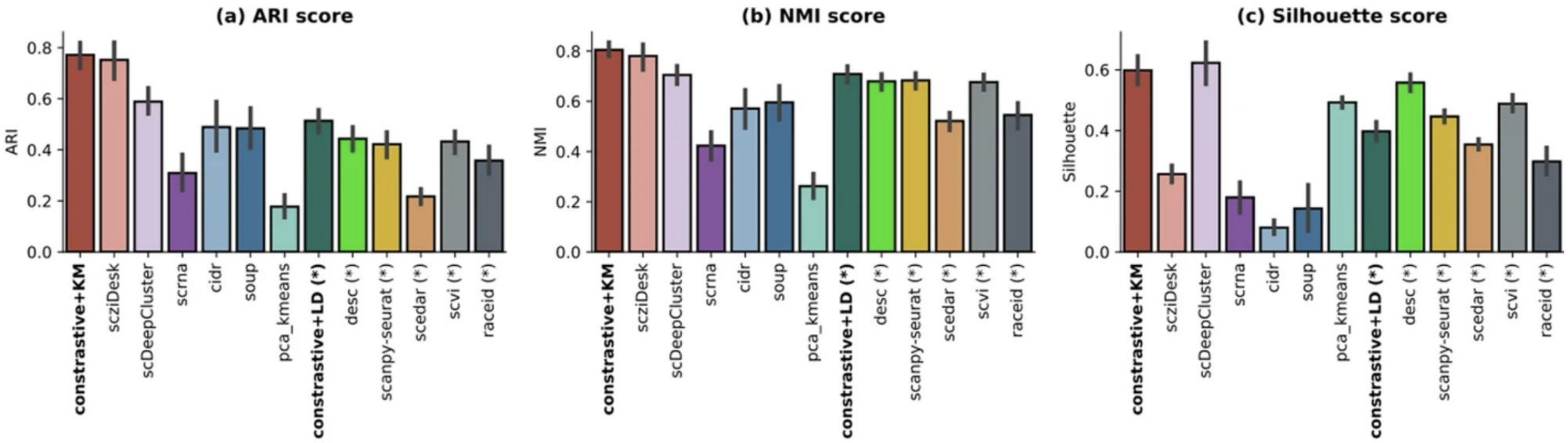
- **Methods:** Contrastive learning (e.g., SimCLR, SupCon), autoencoders, image embeddings, hierarchical clustering, CLIP-style models (optional), dimensionality reduction.

- **Resources for Getting Started**

- **Organizer/Contact**: Chris Endemann (endemann@wisc.edu)



Contrastive Learning Has Been Shown To Improve Clustering Results - Evaluation on 15 real datasets
Autoencoders also work well (scDeepcluster and scziDesk)



[**Ciortan, M., & Defrance, M.** Contrastive Self-supervised Clustering of scRNA-seq Data. *BMC Bioinformatics*, 2021.]

Slides: go.wisc.edu/2et35u

Store Sales: Time Series Forecasting

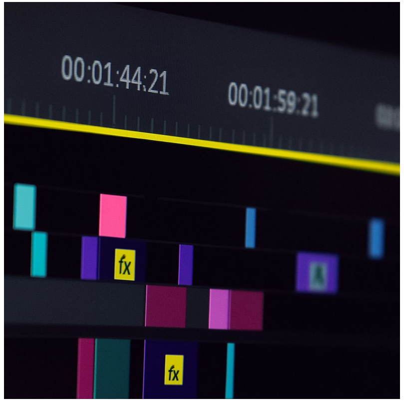
- **Goal**: This challenge introduces core forecasting techniques through a real-world dataset: predicting future sales across multiple retail stores. Participants will explore temporal patterns, seasonality, and exogenous variables to build forecasting models that generalize well over time.

- **Prerequisites**: Basic ML and deep learning foundations are helpful here. No experience with time-series modeling is assumed.

- **Methods:** Classical forecasting (e.g., ARIMA), machine learning approaches (e.g., XGBoost), and deep learning methods for sequence prediction. Optional exploration of holiday effects and promotions.

- **Resources for Getting Started**
    - https://www.kaggle.com/learn/time-series

- **Organizer/Contact**: NA (hosted by Kaggle). See Discussion tab from challenge page to get help.



Slides: go.wisc.edu/2et35u

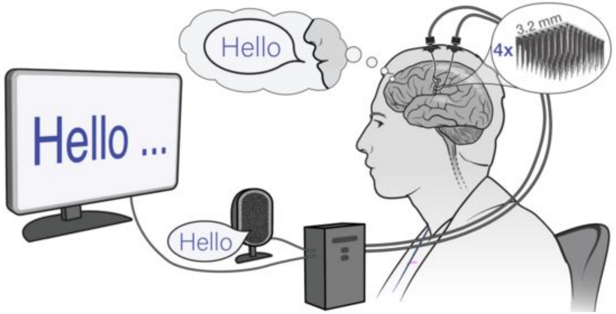
Brain-to-Text '25: Decoding Neural Signals to Speech

- **Goal**: Foster the development of new and improved algorithms for decoding speech from brain activity. Improved accuracies will make it more likely that a speech brain computer interface (BCI) can be clinically translated, improving the lives of those with vocal tract paralysis

- **Prerequisites**: Foundational knowledge of deep learning is expected. Some familiarity with signal processing or time-series modeling is recommended. No neuroscience background is required.

- **Methods:** Signal processing, time-series modeling, spiking neural data, sequence-to-sequence learning, speech decoding, RNNs or transformers, word error rate (WER) evaluation.

- **Organizer/Contact (external)**: https://www.kaggle.com/notnickc
  - Also see Discussion tab on Kaggle



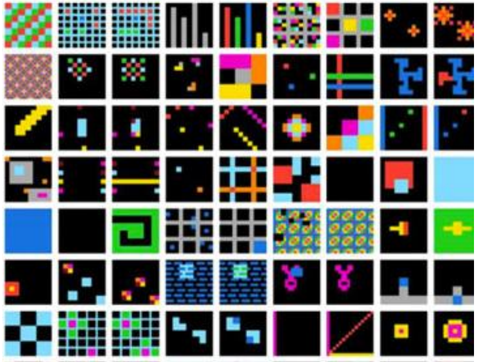
Advanced Reasoning Corpus (ARC) 2025

- **Goal**: Build AI systems that can learn and apply abstract visual reasoning rules from just a few examples, in order to generalize to completely novel problems. Every test task has rules not seen in training. Foundation models excel at interpolation within their training distribution, but ARC is intentionally out-of-distribution.

- **Prerequisites**: <u>For highly advanced participants only</u>. This challenge assumes a high degree of fluency with deep learning frameworks, foundation models, and programmatic reasoning techniques. Participants should be confident exploring advanced, open-ended modeling workflows with minimal guidance.

- **Methods:** Vision Transformers, Few-shot learning, CoT prompting, Reinforcement learning, Pattern transformation logic, Algorithmic reasoning

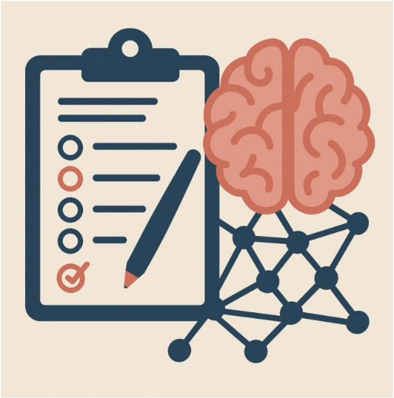
- **Resources for Getting Started**: None. You will be expected to conduct your own research to figure out solutions.



Survey Responder

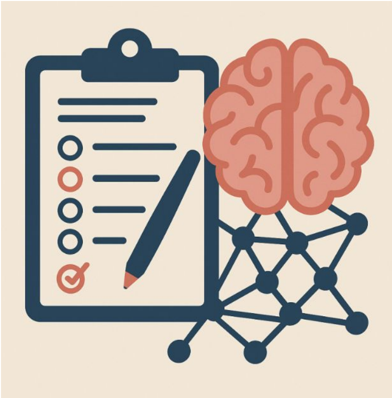
- **Goal**: Investigate if the output from one LLM model may differ from the output of another LLM chat model.
- **Prerequisites**: Knowledge or experience in core machine learning methods (Knn, SkLearn, predictive analytics generally) is optional. The core analytical question for this project is descriptive / or inferential.

Also probably helpful will be: Familiarity with LLMs (e.g., via Ollama or AnywhereLLM) and basic survey research or psychometric principles is recommended.



Survey Responder

- **Methods:** Inferential statistics (t-tests, or regression with binary independent variables).
- **Resources for Getting Started:**
  - The GitHub Repository
  - Previous Response Results
  - Supplemental Project Description
- **Organizer/Contact**: Adam Ross Nelson (UW's Data Science for Human Behavior, and the Wisconsin Undergraduate Research in Data Science Programs).



WISCURDS: dsi.wisc.edu/wiscurds

Questions?

Slides: go.wisc.edu/2et35u

Networking Time – Find Your Team :)

1. **Find signs around the room indicating different challenges**

2. **Share with others…**
    a. Your goal(s) for participating in MLM25
    b. Why you're interested in your selected challenge
    c. Your previous experience working on ML projects
    d. How many hours/week you realistically expect to commit
    e. Anything else you'd like to share

2. **Team registration due by 9/18, but aim to complete it tonight**
    a. https://forms.gle/UrmRK7q3CZjuqUZz6 –
    Only ONE person from each team should fill out the form

    b. **Incomplete teams should also register if struggling to find members** – we'll share a spreadsheet of *incomplete* teams to Slack by Monday

Slides: go.wisc.edu/2et35u

ML+X Nexus: Crowdsourced Resources

go.wisc.edu/68h67v

- Community GitHub repo and website for community members to share and contribute ML resources

- **Resource** = anything that can make ML practice more connected, accessible, efficient, and reproducible
  - External or original

- Includes educational materials, applications & stories, ML tools, and more

ML+X

Slides: go.wisc.edu/2et35u

ML+X Nexus: Crowdsourced Resources

go.wisc.edu/68h67v

How to contribute?

1. Create an issue announcing your plan to add a resource — see ML+X Nexus Issues
2. Fork the ML+X-Nexus repository
3. Clone the forked repository onto your local machine
4. Create a new branch
5. Write the post, commit and push the changes
6. Make a pull (merge) request

Link to slides: go.wisc.edu/52ya98

ML+X

Slides: go.wisc.edu/2et35u

Resources for Getting Started: Kaggle Notebooks

See Kaggle's documentation for more information: kaggle.com/docs/notebooks

- **View your Kaggle notebooks**
  - Code → Your Work
- **Open a notebook**
  - Code → New Notebook
- **Upload data to notebook**
  - Add Data → Search for your challenge
- **Switch from notebook to script**
  - File → Editor type
- **Download code**
  - Click the vertical three dots and select "Download code"
- **Save notebook often!**
  - Save version → Name_YY-MM-DD_HHMM
- **Share notebook**
  - Share → Add collaborators
- **Can I work on a notebook with someone else live?**
  - No, collaborators won't see updates until you save a new version
  - You can sync kaggle notebooks with Git, however

Cloud Resources

1. **Minimum of $50 in AWS/GCP compute credits per person.** You can apply for more based on project needs.

2. **AWS Workshop on Oct. 9th, 4:30-7:30pm**

3. **GCP Workshop on Oct. 30, 4:30-7:30pm**

4. You'll need to pass a short quiz showing that you are familiar your preferred cloud vendor before gaining access. Workshop materials will be provided.

Other Resources

1. Advisor assigned to each challenge starting next week
2. Full hackathon community (#mlm25): We encourage you to post your questions to our Slack channel, #mlm25 (make sure to join Slack group first)
3. Kaggle/learn: Kaggle provides many tutorials and notebooks on various ML topics
4. ML+Coffee (9-11am; 1145 Discovery Building): Monthly social & coworking event with other ML practitioners. Plus free coffee/tea! Our next session is Oct. 9-11am. Sign up to discuss your hackathon project: https://forms.gle/F5LYqnY5TMiZrzXW6

Slides: go.wisc.edu/2et35u

MLM25 — Looking Ahead

1. **9/18 (Thur), 4:30-6:30pm**: Sprint 1 + EDA & Reproducible ML
    a. **11:59PM: Deadline to form team** (sooner is better)

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

3. **10/2 (Thur)**: **Exploratory data analysis presentations**

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