

DataFirst Fall 2023 Projects

Plenary project presentations

- 18 projects
- About 3 minutes each (+1 minute for any plenary questions)

To be followed by:

Breakout sessions

- Opportunity to further explore projects with mentors
 - For mentors attending in person or on Zoom
- Alternative: DataFirst Piazza (see DataFirst website)

Pyleoclim: A Python Package for the Analysis of Paleoclimate Data

Prof. Deborah Khider and Prof. Julien Emile-Geay



Why Paleoclimate?

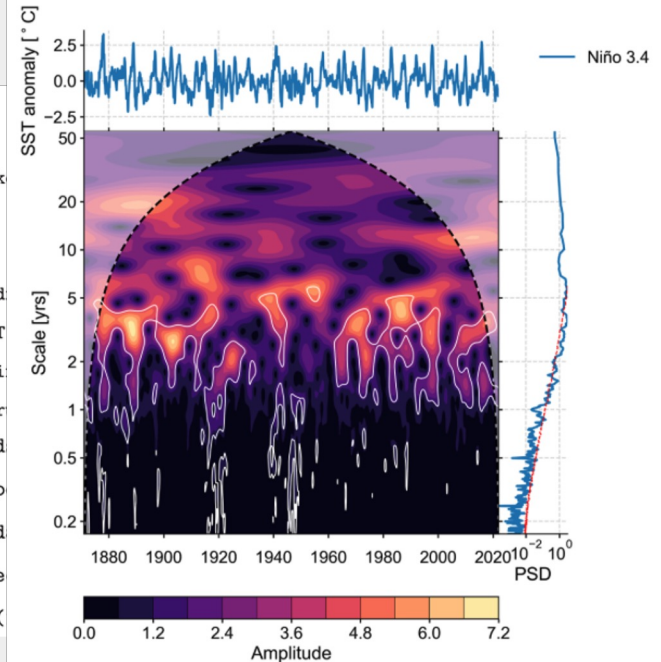
- Understand how climate has changed in the past so we can put the current changes into context.

What is Pyleoclim?

- Python Package for the analysis of paleoclimate time series data
- Automate data transformation
- Automate visualization

A sleek

```
[1] import pandas as pd
[2] import pyleoclim as pyleo
[3] url = 'https://raw.githubusercontent.com/Link
'master/example_data/oni.csv'
[4] df = pd.read_csv(url,header=0)
[5] ts = pyleo.Series(time=df['Dec year'],value=d
        time_name='Year', value_name='SST
        time_unit='CE', value_unit='°\ci
        label='Ni\~{n}o 3.4', clean_ts=Tr
[6] ts_detrend = ts.detrend() # remove trend
[7] ts_interp = ts_detrend.interp() # interp
[8] ts_std = ts_interp.standardize() # stand
[9] PSD = ts_std.spectral(method='mtm') #spe
[10] PSD_signif = PSD.signif_test() #run AR(
```



Pyleoclim: A Python Package for the Analysis of Paleoclimate Data

Prof. Deborah Khider and Prof. Julien Emile-Geay



We have feature requests!

- Visualization, data validation, outlier detection, fluctuation analysis.

How can you help?

1. **Set your time commitment!**
2. Pick your issue based on interest, skills, and time
3. Work on your issue in a Notebook
4. Integrate your functionality into Pyleoclim
5. Write CI tests
6. Write Documentation
7. Write a Tutorial Notebook to teach scientists about your functionality!

Skills needed

Depends on issue

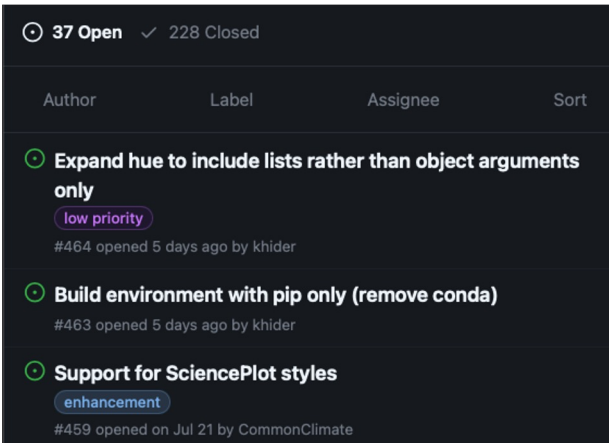
Basic: Python, NumPy, Matplotlib

Advanced: unsupervised learning, scikit-learn, seaborn, pandas

What you will learn

Timeseries analysis, code packaging, project management, GitHub, containerization, building tutorials and documentation, JupyterBook, open source development

<http://linked.earth/PyleoTutorials/intro.html>
khider@usc.edu; julieneg@usc.edu



Learning and forgetting in neural networks

Prof. Marcin Abram

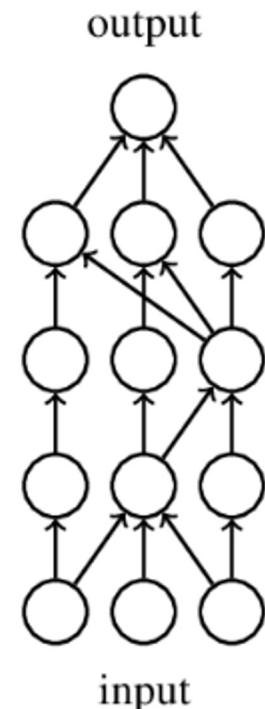
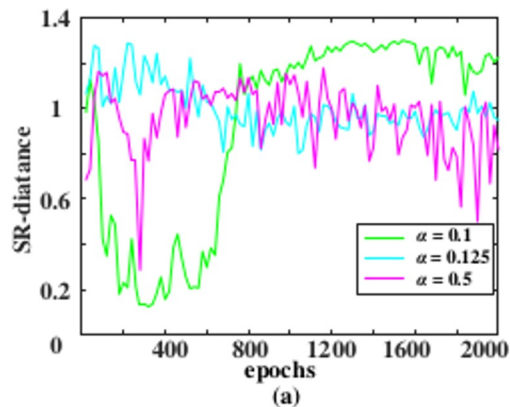
Task: You will study mechanism responsible for **learning** and **forgetting** in artificial neural networks.

Approach: You will investigate **neuron activation patterns** emerging under different conditions (e.g. when learning examples from different data distributions).

Personal Goal: You will **learn** how the information is **acquired** by and **stored** in neural networks.

Scientific Goal: You will explore connections between **machine learning** and **physics** (a sign of potential phase transition).

Skills needed: Python, TensorFlow or PyTorch



Utilizing AI Generated Images for Object Detection and Classification

Prof. Seon Ho Kim

Description: Developing image-based object detection and classification models requires significant time, resources, and effort. Especially, acquiring a good training dataset is essential. However, there are some cases when it is very hard to get quality data such as rare cases (e.g., disasters) or expensive cases to get (e.g., faraway places). Due to the development of generative AI, we might produce synthetic images to enhance the quality of dataset by filling up missing images with them. Based on our prior work in object detection and classification for smart city applications, we would like to explore the potential of AI generated images for an enhanced object detection and classification.

Skills needed: Python, Some knowledge in image machine learning, Knowing existing methods such as Yolo is desirable, Imagination!!!

Students will learn: Image machine learning, object detection applications

Description: Cities are the focal point of economic, social, and environmental challenges and opportunities. To establish USC as a thought leader and partner of choice to tackle the challenges of the urban future, the USC Sol Price School of Public Policy and the USC Marshall School of Business propose establishing an Urban Futures Data Core to serve as a university-wide hub for data analysis and dissemination. Students working on this project will first build an algorithm that gathers information from published research papers by Price and Marshall faculty.

Information collected include dataset name, availability, and geographic level of granularity. Students will also have the opportunity to interview faculty to understand data use agreements. Next, they will build and design a website to present a detailed data catalogue that can be updated in real-time. Finally, students will have the opportunity to use the data to create geographic visualizations of key indices related to urban futures.

Skills needed: Python; web design, web scraping

Students will learn: The students will learn about all data sources used in public policy and business, data management, and web design.



FOCUS AREA	DATA EXAMPLES
1. Real estate	Neighborhood data for social change
2. Housing	House sale price data
3. Infrastructure	Travel diary data
4. Mobility	Cell phone data
5. Energy	Energy consumption data
6. Sustainability	CO2 emissions data
7. Health	Life expectancy, health spending data
8. Skills and jobs	Employment characteristics

Schaeffer Center Data Core

Home Data Sets Code DUAs Sub-Projects Grants IRB Publications Events People Metrics About

🔍

Schaeffer Data Catalog

The Schaeffer Center offers a wide variety of useful resources and tools for researchers in the areas of health policy and economics. Our 60+ data holdings cover a range of data sources including survey data, public and private claims, and electronic health records. The following list does not encompass **all of the data** used by researchers at the Schaeffer Center, only those that are hosted in our analytic computing environment.

Abbreviation	Title ^	Protectio Status
ABC	Abecedarian Limited Data Set	Protected
APCD	All Payer Claims Database	Publicly available
ACS	American Community Survey - Public Use Microdata	Publicly available
ATUS	American Time Use Survey	Publicly available
ARF/AHRF	Area Health Resource File	Publicly available
BRFSS	Behavioral Risk Factor Surveillance System	Publicly available
Open Payments	CMS Complete Program Year Open Payments Dataset	Publicly available

Browse by Data Status

- Schaeffer Data Holdings
- All Data Used at Schaeffer
- Public Use Data Holdings
- Protected Data Holdings
- Data Use Agreements

Browse by Data Type

- Census
- Insurance Claims
- Price Reports & Indices
- Provider & Facility-based
- Support & Reference
- Surveys

Health

> Pharmacies

max: 11.00

min: 0.00

- Above 8.00
- 5.00 - 8.00
- 3.00 - 5.00
- 1.00 - 3.00
- Below 1.00

Map labels: Santa Monica, Culver City, View Park-Windsor Hills, Inglewood, Lennox, Los Angeles International Airport, San Diego Fwy N, 90

AI/ML assisted fault detection in foundry processed devices

Prof. Andrew Rittenbach and Prof. JP Walters

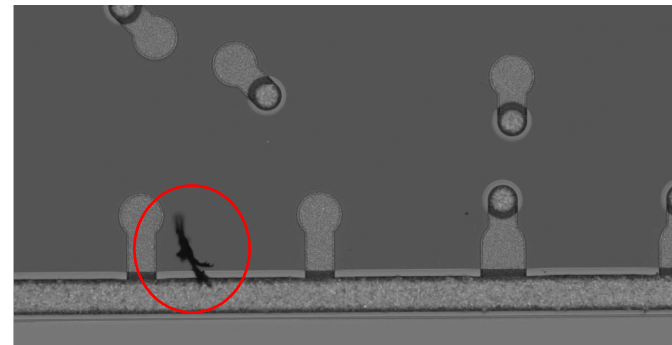
Description: Highly accurate fault detection in foundry produced microelectronics is crucial to ensuring quality of devices that leave the foundry. However, current defect detection flows are human-centric, which produces a bottleneck. The objective of this project is to leverage recent advances in AI/ML to develop automated techniques that can 1) identify manufacturing defects in microelectronics using imagery collected at the foundry, and 2) determine whether the identified defect will impact the performance of the manufactured component.

Skills needed: Python, PyTorch or TensorFlow, image analysis

Students will learn: Students will learn about manufacturing defect detection algorithms, machine learning techniques, and microelectronics fabrication.

AI/ML assisted fault detection in foundry processed devices

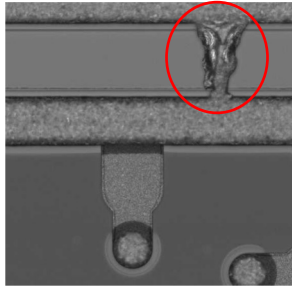
- Highly accurate fault detection in foundry produced microelectronics is crucial to ensuring quality of devices that leave the foundry
- However, many current IC defect detection flows are human-centric and have potential to be a **bottleneck** in the foundry
- Objective of this study is to find ways to leverage recent advances in AI/ML to **enhance and accelerate** the fault detection flow



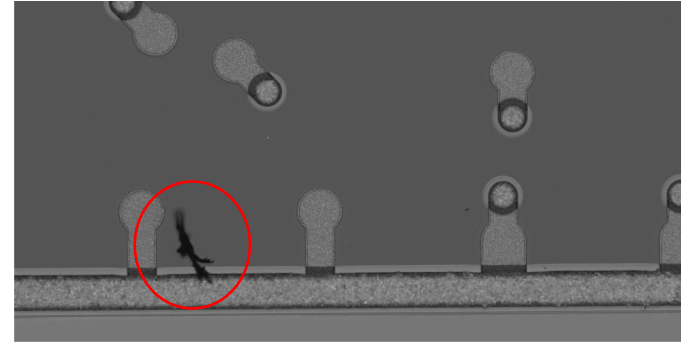
Sample defect- foreign material in IC spanning multiple structures



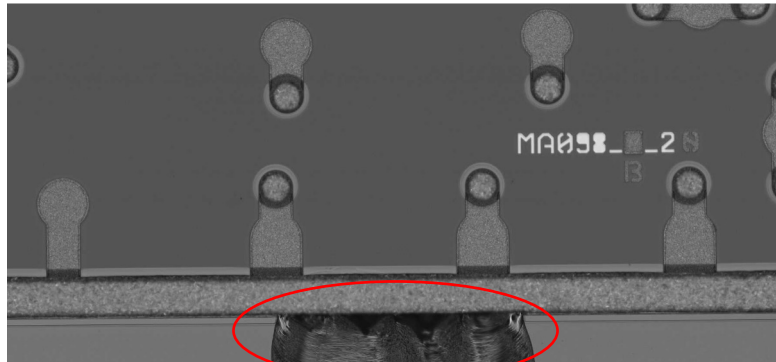
Challenge 1: Many different types of defects



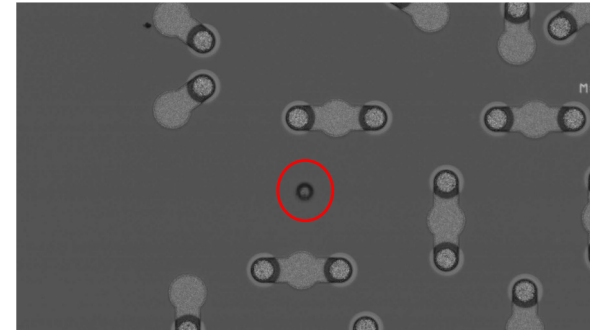
Metal bridging



Foreign material



Chip out

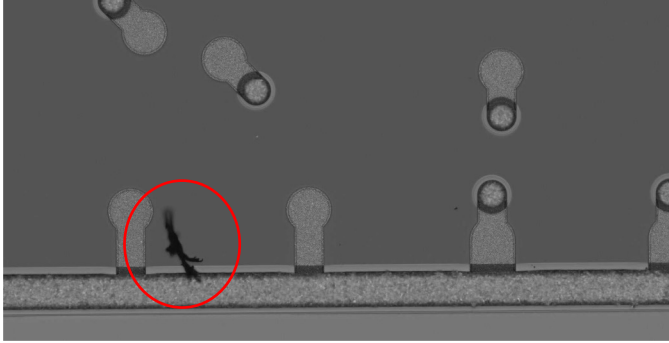


Wafer pitting

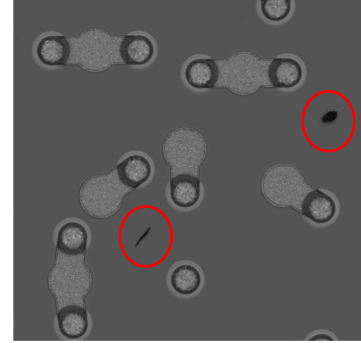
Defects have many different shapes and sizes



Challenge 2: Not all defects are show-stoppers



Foreign material crossing boundaries



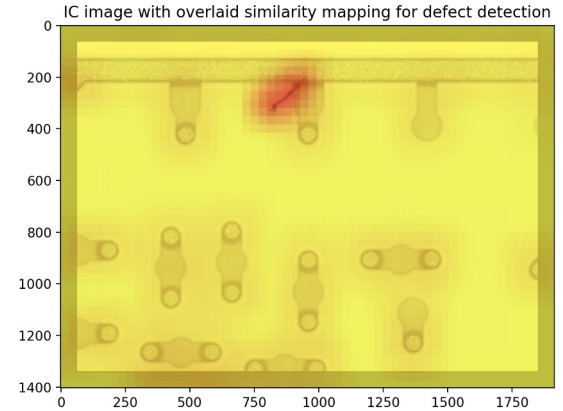
Foreign material in IC wafer

Not all defects impact performance of IC, and
some are considered acceptable



Project objectives

- Due to **large variety of types of defects**, training a model to identify defects using a completely supervised based approach is **not viable**
 - **Difficult to collect large training corpus** with many samples of each type of defect
 - Want model to be able to **identify when IC has a defect even when defect example** was not included in data used to train model
- Better approach is to train model that **learns what IC should look like**, and to recognize when there is a defect
 - Involves development of anomaly detection models
 - Stretch goal is to find a solution that makes use of **pretrained feature extractors**



Sample similarity mapping generated using deep image embeddings for defect detection



Thank you!

If there are any questions, or if you have further interest, please reach out to:

- Andy Rittenbach: arittenb@isi.edu
- JP Walters: jwalters@isi.edu



California Public Sector (CAPS) Job Market: Online Dashboard

Prof. William Resh (Civic Leadership Education and Research (CLEAR) Initiative)

Description: Public sector institutions at local, state, and federal levels are facing an unprecedented hiring crisis in competition for new talent. Yet there is no systematic understanding of the needs and openings across these levels of government to inform stakeholders such as universities, community colleges, and high schools on the current and emerging hiring trends in what constitutes approximately 15-20% of the entire labor market. In this project, students will develop algorithms that continuously scrap relevant job sites used by these governments to assess both developed and emerging hiring trends by aptitudes, professions, entry-levels, mobility, location, and other important attributes. In so doing, the project will inform researchers in public policy, public administration, political science, and labor economics as well as practitioners in government and associated stakeholders.

Skills needed: Python, Statistics, R or Stata, Web Development

Students will learn: Students will learn how to develop and organize labor market data to be used by practitioners and researchers through the construction of portal that can ably transform data into usable aggregated statistics and graphs.

California Public Sector (CAPS) Job Market: Online Dashboard

Prof. William Resh (Civic Leadership Education and Research (CLEAR) Initiative)

(Initial) Purpose of the project

- Systematically collect data that provides insights on the dynamics of the public sector job market within California
 - particular focus on the Los Angeles metropolitan region
 - merge with existing database on federal jobs in CA
 - scrape job listings from all local and state sources in CA
 - Identify and improve existing knowledge of workforce trends for public sector organizations (PSOs)
 - eg., hiring processes; emergent knowledge, skills, and assets (KSAs); needs by regions, positions, levels of government, agencies
- Skills needed: Python and data management**

California Public Sector (CAPS) Job Market: Online Dashboard

Prof. William Resh (Civic Leadership Education and Research (CLEAR) Initiative)

Informative existing projects...



Talent-to-Industry Exchange
PUBLIC SECTOR

A Labor Analysis of Public Sector Occupations In Greater Kansas City

MID-AMERICA REGIONAL COUNCIL
APRIL 2021



Research Report

CHARLES A. GOLDMAN, TERRING DONGADO, J. LUKE BINK, DANIEL ALLEN, YING ZHOU, SUNG HO BYE

Recruiting and Hiring a Diverse and Talented Public-Sector Workforce



Supported by the Workforce Alliance and Southern California Association of Governments



California Public Sector (CAPS) Job Market: Online Dashboard

Prof. William Resh (Civic Leadership Education and Research (CLEAR) Initiative)

Extensions of Purpose

- **Set up the data so that it's accessible to the general public (API project)**
 - Personnel managers at public agencies
 - Potential applicants and incumbents within agencies across various regions
 - University, community college, and high school stakeholders in career development
- **Create a dashboard that provides up-to-date heatmaps with a variety of regional indicators, accessible aggregated statistics, and individual-level job listings**
 - KSA demands
 - Intensity of difficulty in hiring across regions
 - E.g., time listings stay open, salary range vs location cost-of-living
 - KSA supply (by location/region)
 - Competition w/ other sectors
 - Percent of STEM jobs opened per region

The analyses and methods used to build the **API** can be replicated / extended to other metropolitan regions as well
(eg., **Los Angeles vs. San Francisco vs. San Diego vs. ...**)

Description: Build a knowledge graph of a crowdsourcing event.

Skills needed: NEO4j, or Python

Students will learn: Crowdsourcing, knowledge graphs and text analysis

Mentor bio: Prof. O'Leary is Professor of Accounting in the USC Marshall School of Business, focusing on artificial intelligence, emerging technologies and text mining.

Does Municipal Broadband Deliver as Promised? Prof. Hernan Galperin

An examination of broadband pricing and household adoption in areas served by muni networks.

Description:

- Broadband networks owned and/or operated by local governments ("muni networks") are increasingly seen as a key tool to close the digital divide in Internet availability and adoption.
- Only anecdotal evidence about whether muni networks deliver more affordable broadband in communities of limited interest to traditional ISPs - typically disadvantaged communities.
- Goal: to examine broadband pricing and adoption at the address level in areas served by muni networks, using a matched sample of comparable areas as a reference point.
- The project builds on Datafest 2022 project (25K addresses in LA).

Skills needed: Data scraping (Python), statistics and basic GIS skills

Students will learn: Students will have the opportunity to apply data scraping, organization and analysis skills in the context of policy analysis

Automated question type coding of forensic interviews

Prof. Thomas D. Lyon

Question type coding is a method to assess the quality of forensic interviews

- Is this question neutral or leading? Can it be answered by a simple yes or no or is it seeking to elicit more details

When the interviewees are children, such as in cases physical or sexual abuse, question type is particularly important

- Will a 5 year old understand this question?

Most research in the field relies on manual question type coding

- One coder codes every question in the interview, and another coder reliability codes a subsample→ Labour-intensive and time-consuming

Our goal: Develop automated question type coding for forensic interviews with children

Automated question type coding of forensic interviews

Prof. Thomas D. Lyon

The project so far

- Working with a graduate of the Masters in Computer Science program
- Various Large Language Models have been tried - RoBERTa performed the best
- Currently integrating the large language model into Google Sheets to allow for automatic calculations of the quality of the interview

Next steps

- Train the model to make finer distinctions -
- Explore zero shot and few shot prompting to make distinctions for which there is limited manually-coded data

What we need

- Python libraries (preferred)- pandas, huggingface, pytorch OR any coding language

What you'll learn

- Data preprocessing
- Finetuning large language models
- Finding optimal hyperparameter using cross validation techniques
- How to deploy the fine-tuned model on a cloud platform
- Research on zero shot approaches and evaluate its performance with the fine-tuned models

Description: Open source sports data such as the nflverse has led to a massive increase in public sports analytics. But it's still hard to process, subset, visualize and analyze this data. This project will build a general-purpose analysis platform and dashboard, similar to what many teams use internally. Using the nflfastr data, this platform will allow interested individuals to select the play parameters they're interested in, and will provide relevant analysis, visualization and insight. Ideally, we'll set up the dashboard on the internet, and open source the project, allowing others to expand the available datasets, analyses and visualizations.















Skills needed: Python (pandas, streamlit, dash, etc. is a bonus!)

Students will learn: How to analyze and present insights from NFL play-by-play data

Understanding the Relation Between Noise and Bias in Annotated Datasets

Negar Mokhberian

Research question: Is it possible to extract useful signal from annotators' disagreement existing in the training data samples?

	Prompt: Do you see an airplane in the given picture?				Prompt: Do you think the given text is offensive?				
									
	✓	✓	✗	✓		✓	✗	✗	✓
	✗	✗	✗	✗		✗	✓	✗	✗
	✓	✗	✓	✓		✓	✗	✓	✗

Understanding the Relation Between Noise and Bias in Annotated Datasets

Negar Mokhberian

- many previous work has tried to design larger and more complex neural networks.
- data-centric AI has worked on shifting the focus to the quality of the train data.
 - Annotations are can exhibit both noise, stemming from vague instructions or human errors, and bias, arising from differing perspectives among annotators in response to given prompts.
- In this project, our objective is to bridge the gap between the two lines of research: **Our investigation will center on whether perspectivist classification models have effectively harnessed valuable information from instances flagged as noisy by noise-detection techniques.**

What kind of classification task? We will work on the domain of offensive text detection datasets, a highly subjective task.

Skills needed: Python, PyTorch, Fine-tuning language models in Huggingface package

Can we learn from distributed data respecting privacy?



Learn a model over all the data without sharing data?

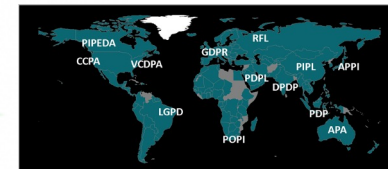
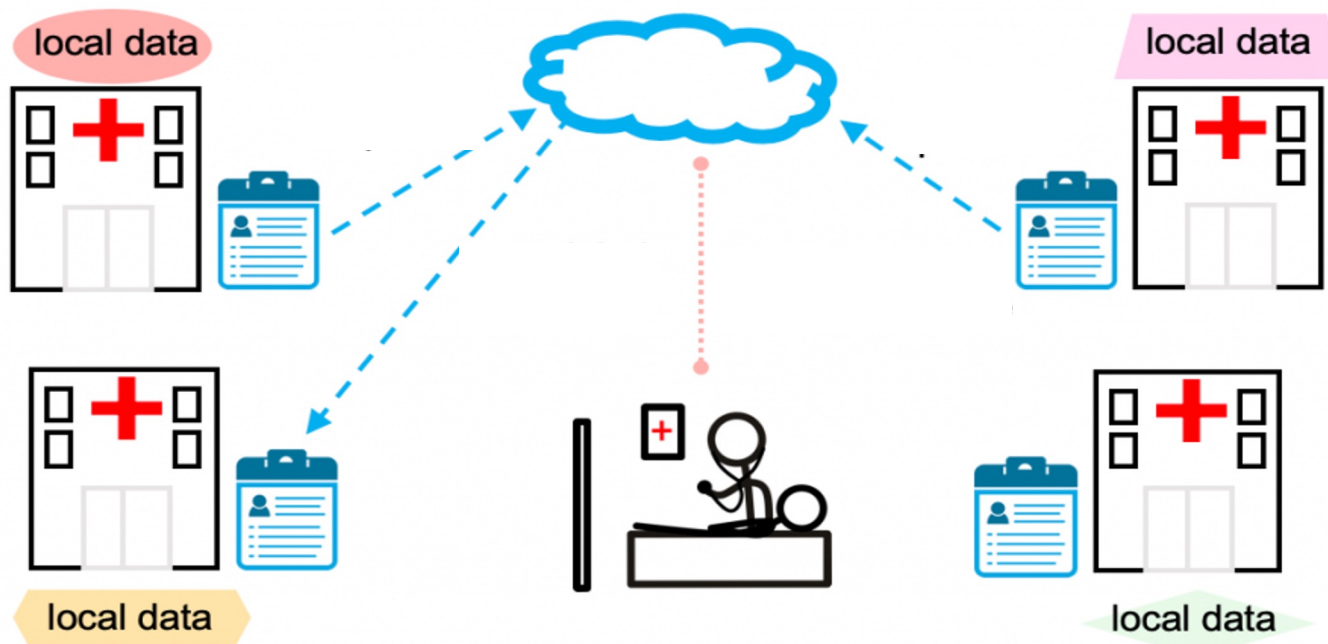
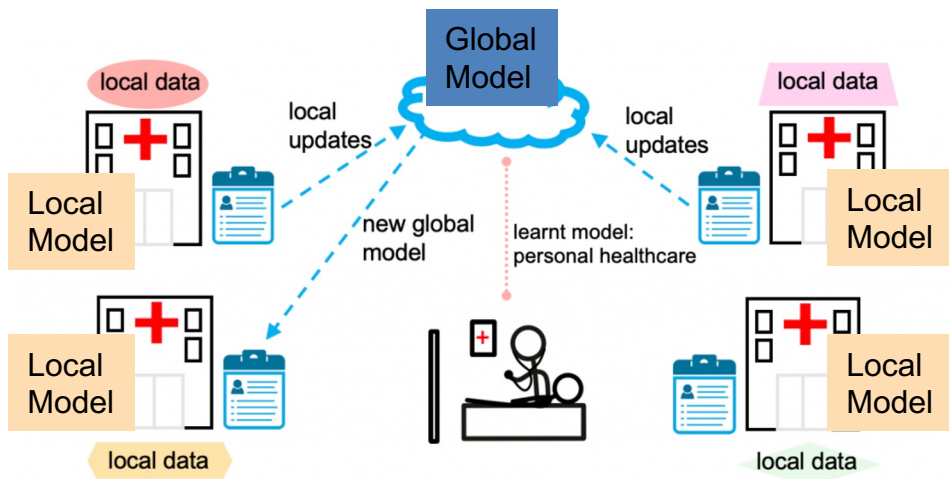


image courtesy: <https://blog.ml.cmu.edu/2019/11/12/federated-learning-challenges-methods-and-future-directions/>

Federated Learning: Learning without sharing data!



- No data shared among data sources
- Training is pushed down to data sources
- Sources share parameters (e.g., weights, gradients), but parameters are strongly protected (encryption + gradient noise)



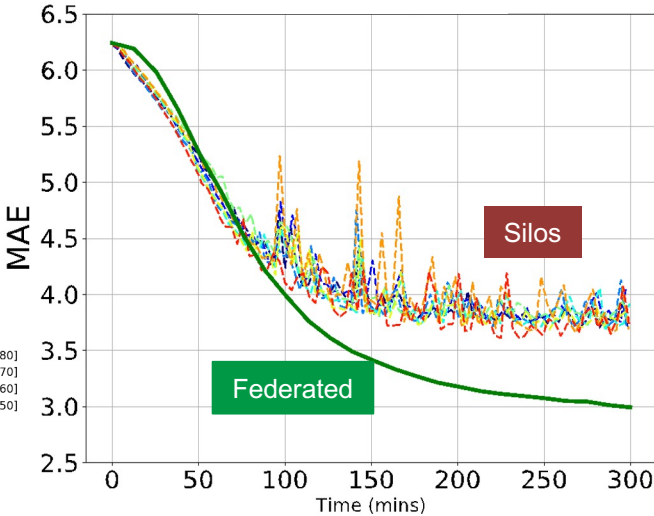
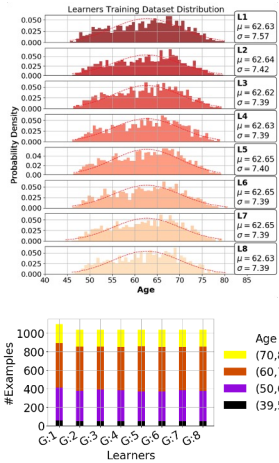


Federated Learning Outperforms any Silo

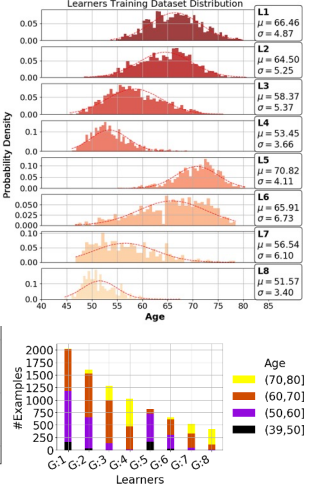
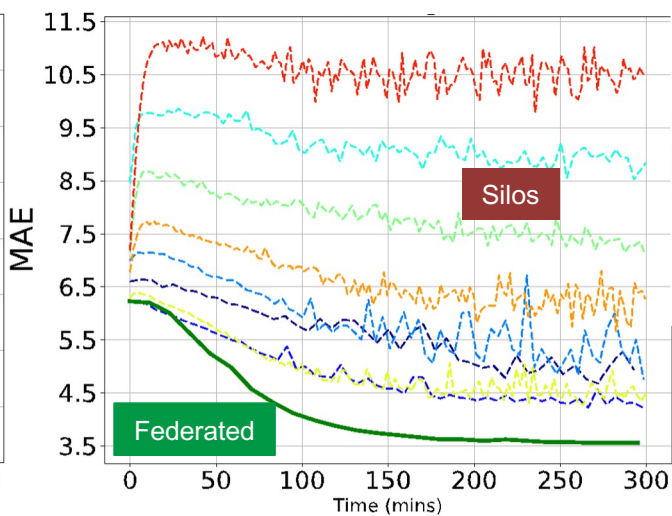
BrainAGE: predict age from MRI scan

Uniform & IID

Skewed & Non-IID



- G:1 - - - G:4 - - - G:7
- G:2 - - - G:5 - - - G:8
- G:3 - - - G:6 - - - Community

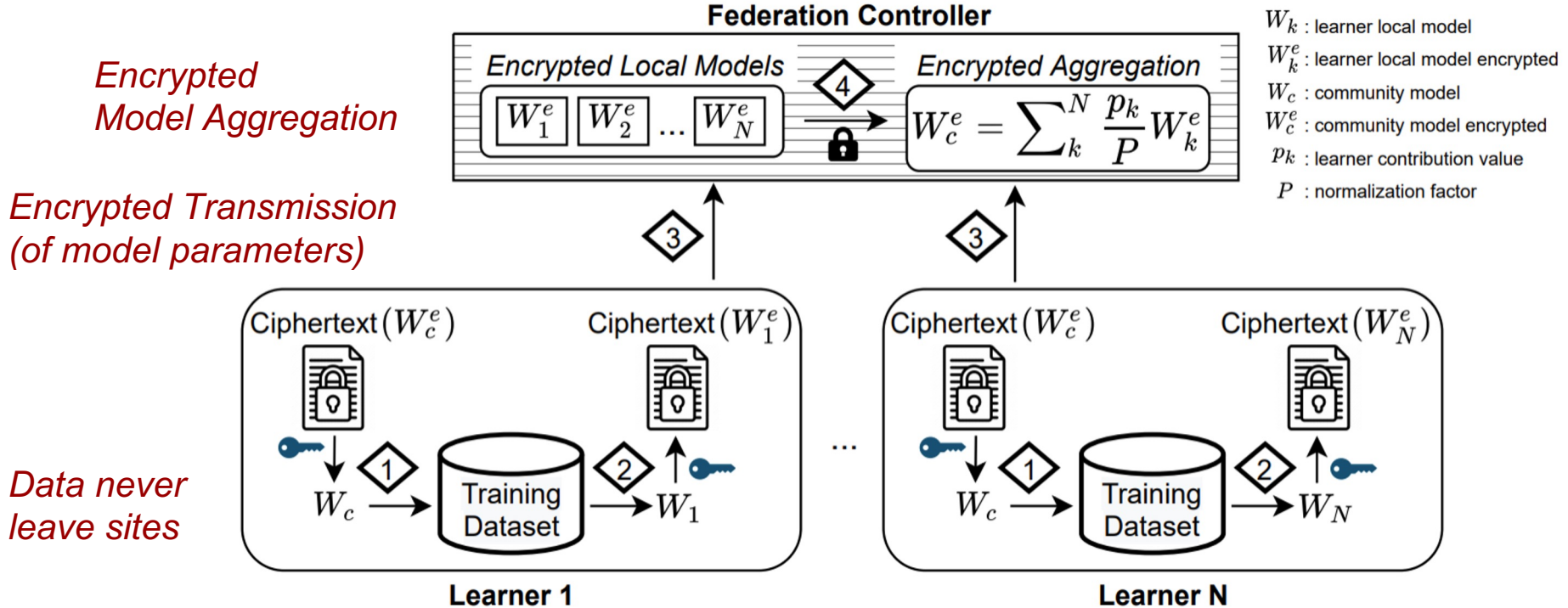


- G:1 - - - G:4 - - - G:7
- G:2 - - - G:5 - - - G:8
- G:3 - - - G:6 - - - Community

All sites incentivized to join the federation!



Secure Federated Learning

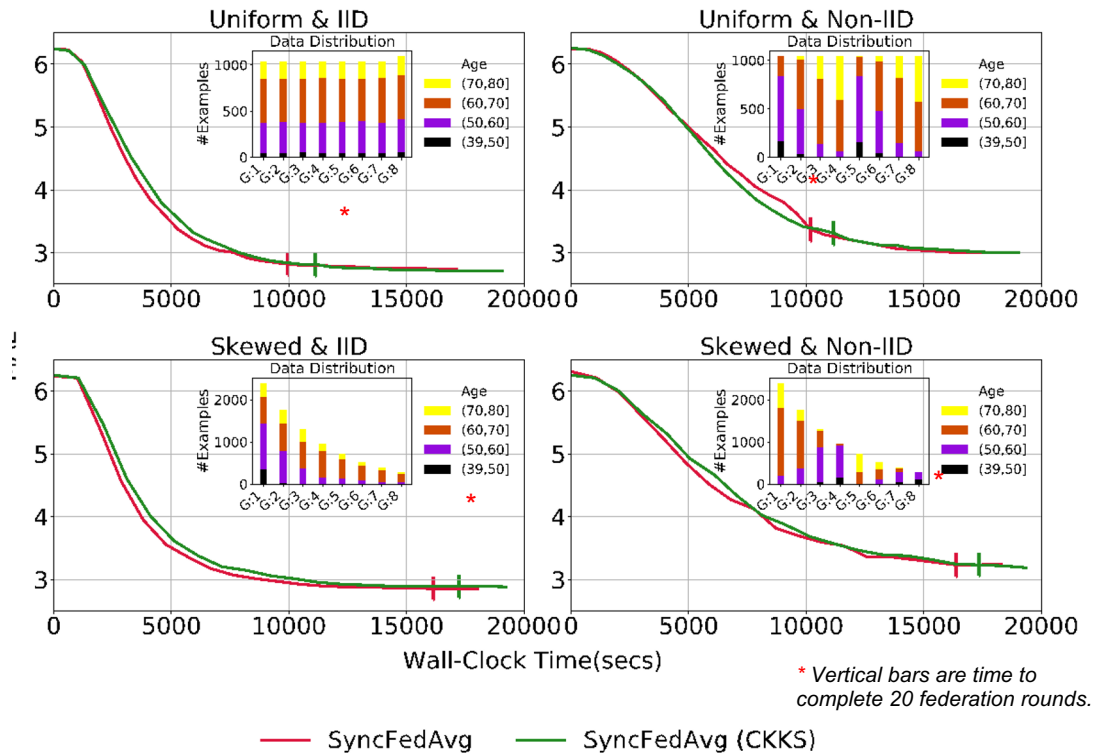


Secure Aggregation with Fully Homomorphic Encryption



- Federated Training with Fully Homomorphic Encryption (FHE) learns model with **same learning performance**
- CKKS FHE (Optimized)
- Low overhead (**~7%**)

BrainAGE (3D-CNN)



Description: Federated learning is an approach to distributed deep learning without sharing data. Multiple sites train a neural network over private data. The parameters of the neural network are shared with a federation controller, but they are encrypted before sharing. Model aggregation is performed under fully homomorphic encryption. We propose to apply federated learning to several problems in neuroscience, such as predicting Alzheimer's, Parkinson's, epilepsy, and autism, possibly over multimodal data.

Skills needed: Python, Tensorflow, deep learning

Students will learn: Federated learning, machine learning for biomedical applications.

Bad Writing is "Fine": Tuning an LLM to Suggest Improvements

Prof. Benjamin Nye

Description: Prototype an approach to fine-tune a large language model (LLM) to help diagnose areas to improve a specific writing product. For example, scientific papers require consistent language but in creative writing variety matters. Proposed steps are: 1. Writing Product: Coordinate with project mentors to choose a common and important writing product, such as a position paper or an academic conference. Identify/gather a rubric and a corpus. 3. Inject Bad Writing: For each element of the rubric, develop prompts for generative AI to decrease the quality of writing based on the rubric (i.e., make it worse). This will form a training data set of the good example and version worse on certain characteristics. 4. Fine Tune: Students will be expected to attempt to fine tune an LLM (e.g., LLAMA 2) based on this synthetically generated data 5. Evaluate: Research if tuning suggests better domain-specific areas to improve. This project aligns with ongoing work with the USC Generative AI Center.

Skills needed: Python

Students will learn: Generative AI for large language models. Generating synthetic data for a rubric. Fine tuning a large language model, likely using CARC (the on campus computing cluster). Understanding intelligent tutoring system design fundamentals for modeling how experts diagnose issues from novices.

Bad Writing is "Fine": Tuning an LLM to Suggest Improvements

Prof. Benjamin Nye

Good Example:

Four score and seven years ago our fathers brought forth, on this continent, a new nation, conceived in Liberty, and dedicated to the proposition that all men are created equal.

Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived, and so dedicated, can long endure. We are met on a great battlefield of that war. We have come to dedicate a portion of that field, as a final resting-place for those who here gave their lives, that that nation might live. It is altogether fitting and proper that we should do this.

But, in a larger sense, we can not dedicate...

Focus	Purpose is clear and presented to the reader <i>Example Issue: Add additional points to the thesis statement that are not necessary to the essay</i>
Content	Well-presented and argued; ideas are detailed, well-developed, supported with specific evidence & facts <i>Example issue: Argument made without support by examples/evidence</i>

(Bad Example, Rubric Issue, Text Diff)

Worse Example:

Four score and seven years ago our fathers brought forth, on this continent, a new nation, conceived in Liberty and dedicated to the proposition that all men are created equal. Additionally, it's important to note that during that time, they were also faced with various economic challenges and social issues that shaped the early development of the nation.

Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived, and so dedicated, can long endure. We are met on a great battlefield of that war. We have come to dedicate a portion of that field, as a final resting-place for those who here gave their lives, that that nation might live. It is altogether fitting and proper that we should do this.

Analyzing Open Source Software Ecosystems

Dr. Jeremy Abramson, Dr. Jim Blythe, Dr. Alexey Tregubov

Description: Open source runs a lot of the world's critical software systems, but there is much that's unknown in how maintainers, developers and other parts of the software ecosystem function. Help us analyze a large corpus of open source data — both source code and patch conversations — to better understand them! We'll study things like rise to influence, authorship styles, malware analysis, topic modeling and social network analysis!

Skills needed: Python (needed), experience with LLMs/OpenAI APIs, program analysis, C code (preferred, but not necessary!)

Students will learn: We'll touch on using LLMs to parse text messages and analyze code, graph databases, program analysis, and social network analysis among other skill

Analyzing Open Source Software (OSS) Ecosystems



Goals:

- Analyze a large corpus of open source data (Linux kernel mailing list) — both source code and patch conversations — to better understand them.
- Help protect the health of OSS code and communities



Project inputs:

- 4.2M LKML messages



Project outputs:

- Extracted structured code and related patch conversations, labeled with **topics, conversation summary, acceptance/rejection reasons**.
- High level reasoning about patch acceptance/rejection using LLM (e.g. GPT-4).
- Labeled data for future projects/analysis.

git: dd95b39235dd - main - [PowerPC64] Fix

Brandon Bergren [bdragon at FreeBSD.org](#)

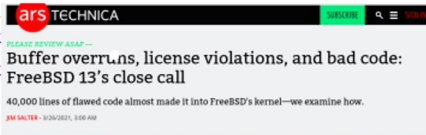
Mon Mar 1 02:38:24 UTC 2021

- Next message (by thread): [git: 384ee7cc6e9e - main - \[PowerPC\] \[PowerPC_SPE\] Fix mult](#)
- Messages sorted by: [\[date \]](#) [\[thread \]](#) [\[subject \]](#) [\[author \]](#)

The branch main has been updated by bdragon:

URL: <https://cgit.FreeBSD.org/src/commit/?id=dd95b39235dd81c890aa3cce02a5bb7f91>

```
commit dd95b39235dd81c890aa3cce02a5bb7f91
Author: Brandon Bergren <bdragon at FreeBSD.org>
AuthorDate: 2021-03-01 02:35:53 +0000
Commit: Brandon Bergren <bdragon at FreeBSD.org>
CommitDate: 2021-03-01 02:37:48 +0000
```



Build a multilingual decipherment system

Prof. Jonathan May

Description: We will build a working system that can decipher a letter substitution cipher into 14 languages and beyond, based on <https://aclanthology.org/2021.acl-long.561/> then apply it to languages it has never seen.

Skills needed: Python, deep learning with transformers

Students will learn: read and understand an NLP paper, unusual applications of transformers, reproduction study

Natural language processing of safety reports in nuclear power plants

Prof. Najmedin Meshkati

Description: Use Natural Language Processing (NLP) techniques to analyze voluminous (30K+ pages) Diablo Canyon Independent Safety Committee (DCISC) annual reports to identify the role and contribution of "Traits of a Healthy Nuclear Safety Culture", as defined by the Nuclear Regulatory Commission and the Institute of Nuclear Power Operations, in incident causation.

Skills needed: Natural Language Processing and related skills.

Students will learn: Application of NLP in real-world, working on very serious and important issues with global applications, which can be generalized and applied to other safety-sensitive technologies.

Application of AI, ML and NLP in understanding and preventing a serious aviation safety problem in the US - Runway Safety Prof. Najmedin Meshkati

Description: Use AI/ML/NLP to understand root-causes of one of the most serious aviation safety problem in the US - runway incursions. The Aviation Safety Reporting Systems (ASRS), which is administered by NASA and is an untapped treasure trove of text data, will be used for this project.

Skills needed: AI/ML/NLP and associated skills.

Students will learn: Using AI/ML/NLP and working on the data from a major global industry - aviation.

AI Ethics for Smart Health through Smart Watches Prof. Yolanda Gil

Description: This project will conduct a thorough study of the ethical issues in using AI systems in this domain, with recommendations of how AI systems for smart health should be designed with ethical considerations in mind.

Skills needed: Interest in AI ethics, healthcare and data science.

Students will learn: What kinds of health-related data can be captured through wearable devices, what kinds of analyses are possible, privacy and ethical aspects of personal applications for smart health.

