WOMEN IN DATA SCIENCE
LIVERMORE

Data Science in Action: Research, Internships, and Mentoring

Nisha Mulakken
Co-Director, Data Science Summer Institute
Biosecurity Bioinformaticist

LLNL-PRES-820231 This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC

Lawrence Livermore National Laboratory

My journey in data science at the Lab

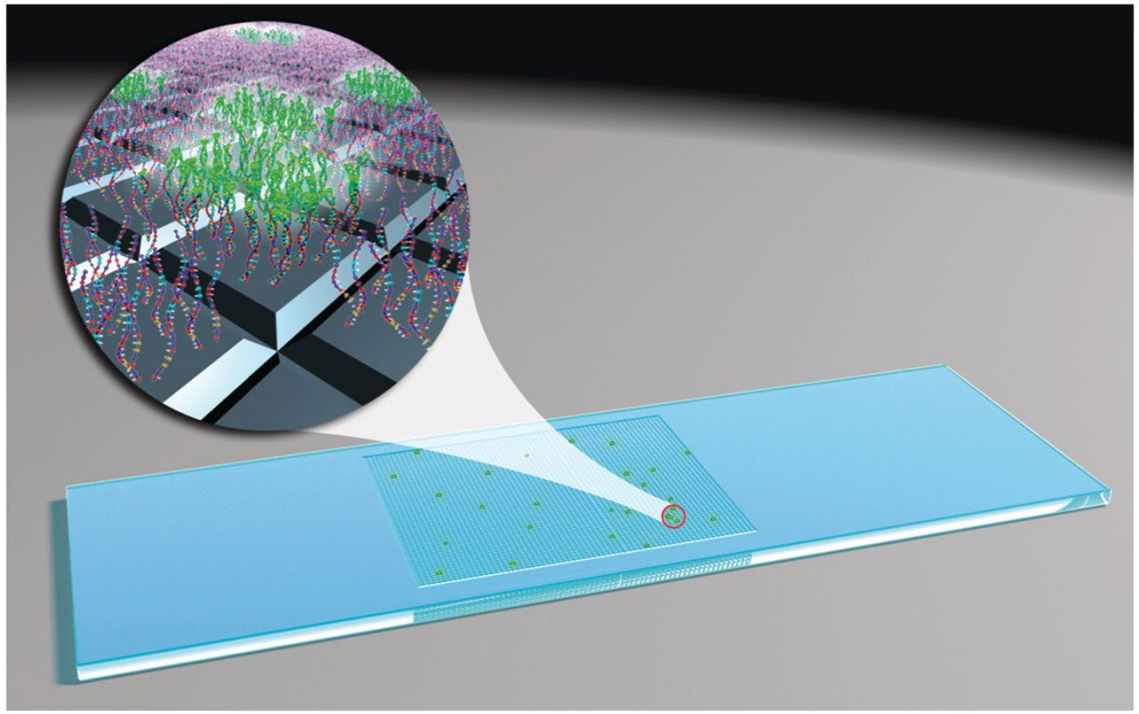
- My biosecurity bioinformatics projects
- Data Science Institute & DSSI internship program
- My summer student's DSSI project
- Mentoring tips

What is biosecurity?

- One of the Lab's missions
- To keep the world safe from ever-changing biological threats
- To safeguard against disease
- Revolutionary advances in detection, characterization, and mitigation

Large-scale pathogen detection and microbiome analysis

- The LLMDA is printed with DNA probes based on all available genomic data for >20,000 species, including viruses, bacteria, archaea, protozoa, and fungi.
- High-performance computing (121,000 cluster CPU-hours) is used to compare DNA sequence regions to find 1.4 million unique signatures to organisms.

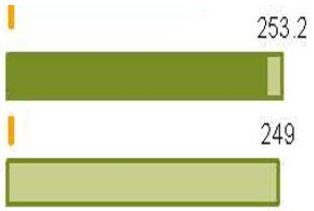


LLMDA analysis: composite likelihood maximization method

- Expected data: modeled by database of probe profiles per organism
  - Probability of each probe to bind to each organism
- Observed data: Calculate the reverse!
  - Probability of organism presence, given observed probes
- Compute composite likelihood function for observed probes
- Rank organisms that best explains pattern of hybridized probes

LLMDA identified *Yersinia pestis* in the tooth of a plague victim from 1348. (Ancient DNA was heavily degraded, with sizes in the 35–50 bp range, making detection by PCR difficult.)

Log odds

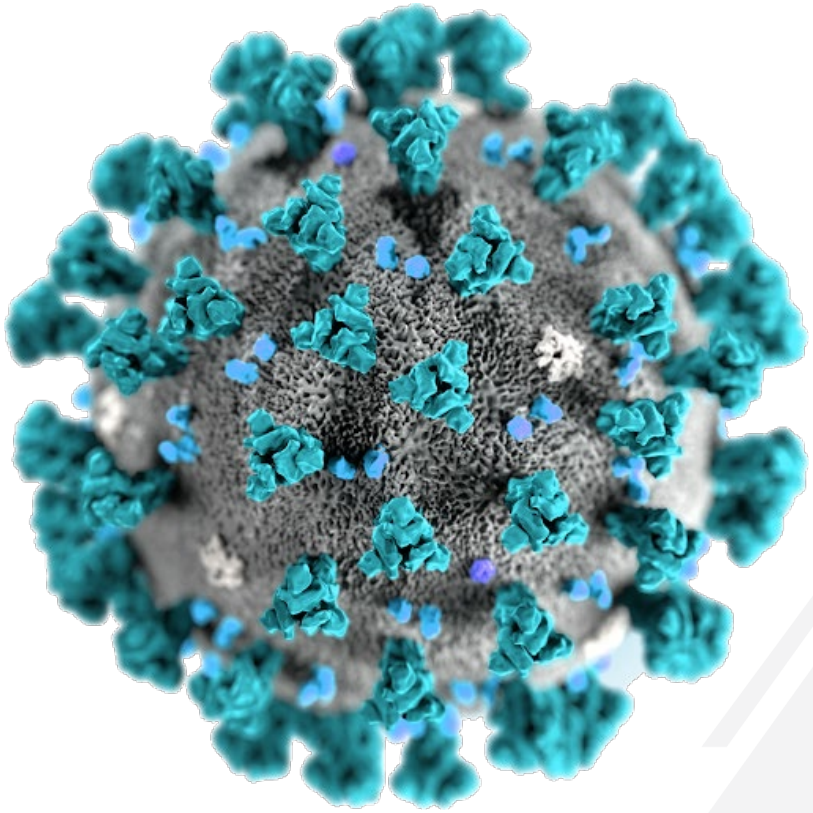


Highest scoring organism

GF:730739 (Yersinia pestis biovar Orientalis str.
MG05-1020 Yersinia pestis biovar Orientalis str.
GF:727233 (Yersinia pestis biovar Orientalis str.
IP275 Yersinia pestis biovar Orientalis str.

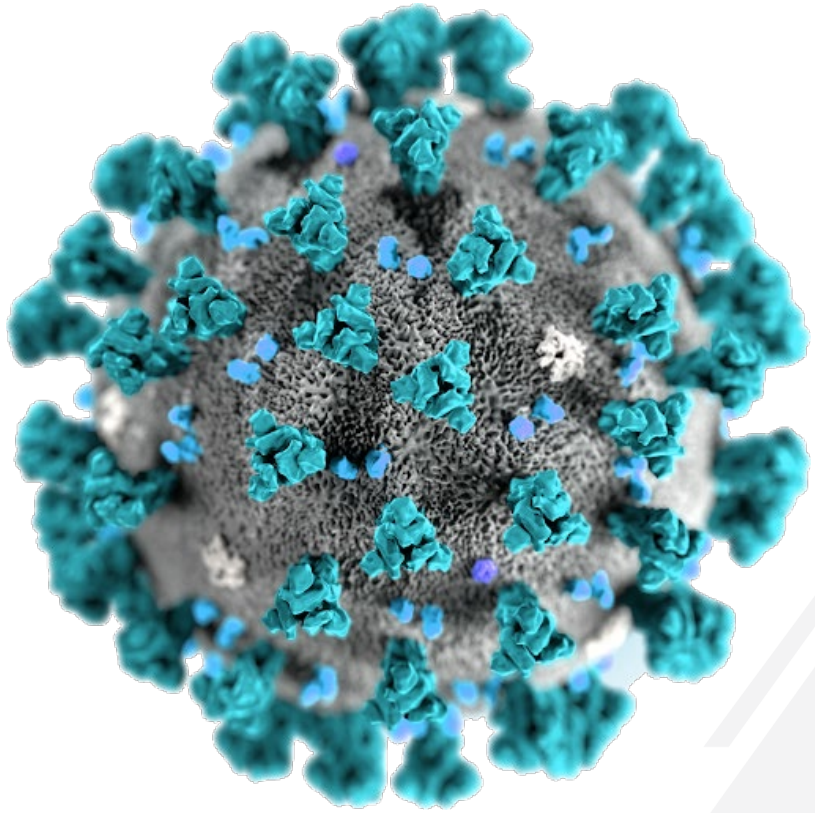
The addition of SARS-CoV-2 probes to LLMDA

- Last summer, >40,000 reference sequences were available from https://www.gisaid.org.
- Selected 78 probes out of >75,000 candidates such that
  - Probes spanned the entire viral genome (this protects against degraded samples)
  - Some probes were conserved across all variants (these are likely to be conserved among future mutant strains)
  - All probes are unique to SARS-CoV-2
- Now >500,000 sequences are available.
- The initial probes can pick up the new variant strains.



Why are some people asymptomatic while others have severe symptoms?

- Currently, this version of the LLMDA is being used to look at SARS-CoV-2 co-infections and metagenomics in patient samples from CA Dept of Public Health.
- Goal: Determine co-infections and microbiome impact on disease severity
- Questions
  - Does the microbiome have a protective effect in asymptomatic and mild cases?
  - Are we seeing multiple strains of SARS-CoV-2 in patients?
  - Do patients with severe symptoms have other known or unknown co-infections?



Addressing the rapid growth of data science and its impact on LLNL's mission

National Security

Cognitive Simulation

Precision Medicine

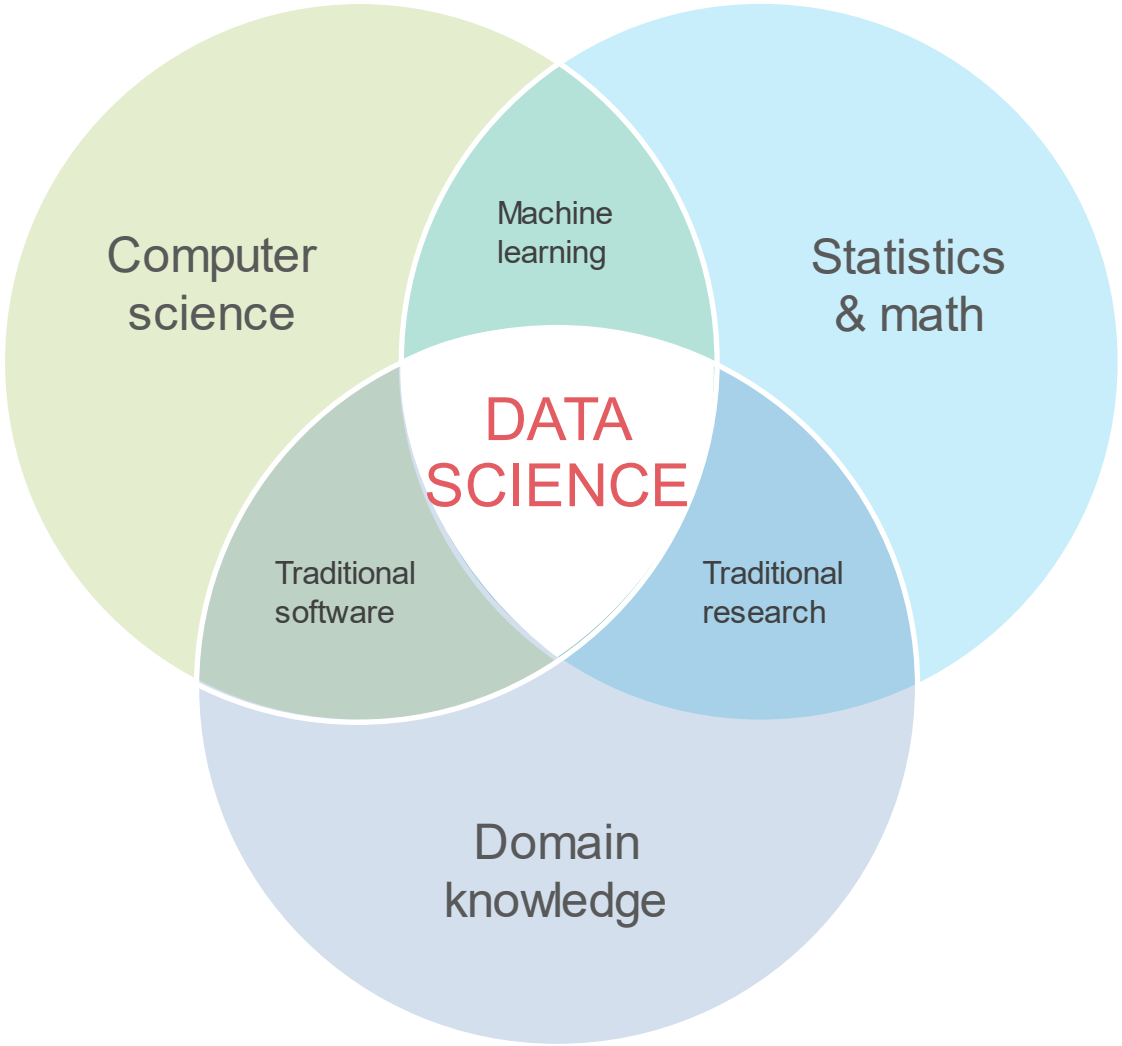
Basic Science

Materials & Advanced Manufacturing

Energy

See our strategic plan for more: **data-science.llnl.gov/about**

Data Science Summer Institute



- 12-week-long internship for undergraduates and graduate students.
- 50% time on project guided by a mentor & 50% on DSSI activities
- Collaborate with other interns on a real-world Challenge Problem
- Data science–related short courses
- Summer SLAM!

Data Science Summer Institute

	FY17	FY18	FY19	FY20
Applicants	100	1,000	1800+	2,000+
Accepted	24	26	32	27
Female Students	7 (29%)	6 (24%)	14 (44%)	8 (30%)
Visiting faculty	Robert Gramacy, Virginia Tech	James Flegal, UC Riverside	Ryan Farrell, BYU Bruce Sanso, UC Santa Cruz	Dorit Hammerling, Colorado School of Mines
Challenge datasets	Topology optimization Cyber security	Machine vision Multimodal physics data	Molecular structures	Human connectome Nanomaterial synthesis

Student project: CRISPR application predictor

Goal

- Classify bioengineered plasmids with their lab-of-origin, recreate (or outperform) results from paper

Potential uses

- Identify potential nefarious actors proactively

Data

- Addgene (public database with data stored in nested JSON format)
- 60,000 bioengineered plasmids
- 1,400 associated lab depositors

Credit: Emilia Grzesiak

Student project: data cleaning

The Challenge

- Paper used old version of Addgene database where Lab depositors shown by Lab name labels
- LLNL database version: Lab depositors had numerical IDs but no “name” labels

The Solution

- Instead of relying on bioinformatics packages that make max of 3 requests/min, web-scraped Addgene website
- Used SLURM arrays to bypass Addgene requests limit → able to make 1,400 requests all at once

Credit: Emilia Grzesiak

Mentoring: lessons learned

Prepare

- Project vision and goals
- Step-by-step tasks by priority
- Set up tools and work environment
- Tips for cleaning data

Guide

- Be available as needed
- Meet daily at least for the first two weeks
- Meet at least once a week after that

Motivate

- Give constructive feedback with compassion and empathy
- Celebrate small victories

Challenge

- Challenge problems
- Provide optional tasks in case project is completed early

Develop Skills

Develop leadership and professional skills:

- Foster independent thinking through autonomy to try out new ideas
- Present project progress to multi-disciplinary team
- Encourage collaboration with peers

Build Career-Long Professional Relationships

Data Science Summer Institute

dssi.llnl.gov
dssi-info@llnl.gov

Lawrence Livermore National Laboratory