IT SERVICES

Data Science/Machine Learning Platform for students

Valerie E. Polichar, Ph.D.
Senior Director, Academic Technology Services

Adam Tilghman
Senior Architect/Analyst, Academic Technology Services

UC San Diego

Inception

Larry Smarr & Thomas DeFanti had created the PRP and are funded for CHASE-CI, which puts their novel-design low-priced, fast GPU cluster, dubbed the “FIONA,” in the hands of researchers, putting a machine learning platform within reach.

In 2017, Larry challenges UC San Diego to do the same thing for their graduate students.

Usual way: Go to a vendor, ask them to spec out a system.

Pros:

- Familiar, “safe”
- Higher level of hardware support
- Possibly software support

Cons

- Very expensive
- Not innovative or cutting-edge — needed technology may not yet be available
- Limited by vendor expertise in novel field
- Fails to leverage campus expertise!

What if we could learn from our world-renowned researchers?

Pros:

- Innovative
- Cutting-edge thinking
- Very efficient
- Access to on-campus expertise
- Ability to share experiences and troubleshoot problems with others

Cons

- “We haven't done this before”
- Design your own
- No “four-hour replacement”
  - Enterprise IT has lower risk but less cutting-edge
  - Research/Instructional IT often has higher risk but greater capacity/more cutting-edge

History

2017:

- CIO Vince Kellen and Sr. Director Valerie Polichar accept the challenge and invest $75K of scrounged core funds to develop an initial system.
- Adam Tilghman architects UC San Diego's DS/MLP inspired by FIONA design specs; team rolls out in Sept. to initial 10 undergrad & grad classes

2019: Added access for student independent study & researchers with entry-level needs

2017-2024: System grows & is supported by annual influx of small dollars saved from other areas to prioritize student learning

As of early 2024: Almost 72,000 student enrollments in 620 classes,
170 instructors in all 9 divisions —>
About ~8-10K unique students per year

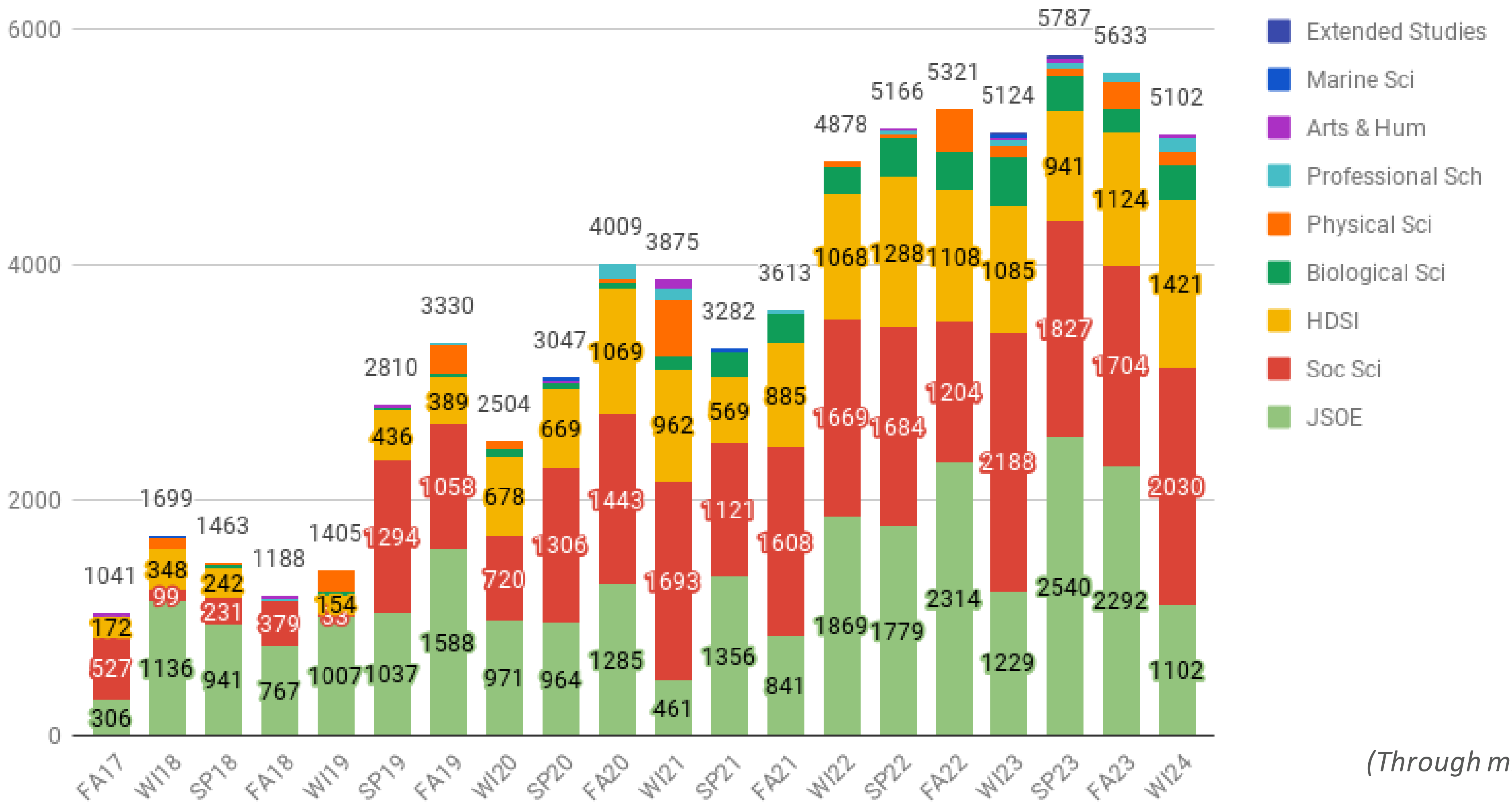
A variety of offerings support many disciplines…

- Jupyter Notebooks
  - Web-based notebooks allow students to combine live code (many languages available), equations, visualizations and narrative text for data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and more.
- Shell accounts
  - Complex ML workflows are supported through terminal/SSH logins, background batch jobs, and a full Linux/Ubuntu CUDA development suite.
- Flexibility
  - Instructors can install additional library packages (e.g. conda/pip, CRAN) as needed, or can opt to replace the default environment entirely by launching their own custom Docker containers.

What this does for our students

- Training available, in every division and many disciplines, in the skills required to perform, design, or build machine learning and/or data science appropriate for that field
- Our students graduate with a competitive background that prepares them for tomorrow’s jobs
- Our students who go on to higher learning do so with a sophisticated toolkit

DSMLP Students by Division, Term



(Through mid-Winter 2024)

Selected Courses (Winter / Spring 2024)

Compositional Algorithms	*Music*	G
Intro to Causal Inference	*Data Sci.*	G
Computational Physics: Probabilistic Models/Sim.	*Physics*	*Mixed*
Environmental Data Science	*Comp&Soc.Sci.*	UG
Machine Learning in Computational Mechanics	*Structural Eng.*	G
Intro to Parallel Computing	*Computer Sci.*	UG
Neural Data Science	*Biology/Neurosci.*	*Mixed*
Machine Learning Competitions	*Data Sci.*	*Mixed*
Political Science Methodology	*Political Sci.*	UG
Statistical Inference in the Medical Sciences	*Biomedical Sci.*	G

Courses have included Visual Arts, Public Health and other “unexpected” disciplines – as well as the nascent Data Science department!

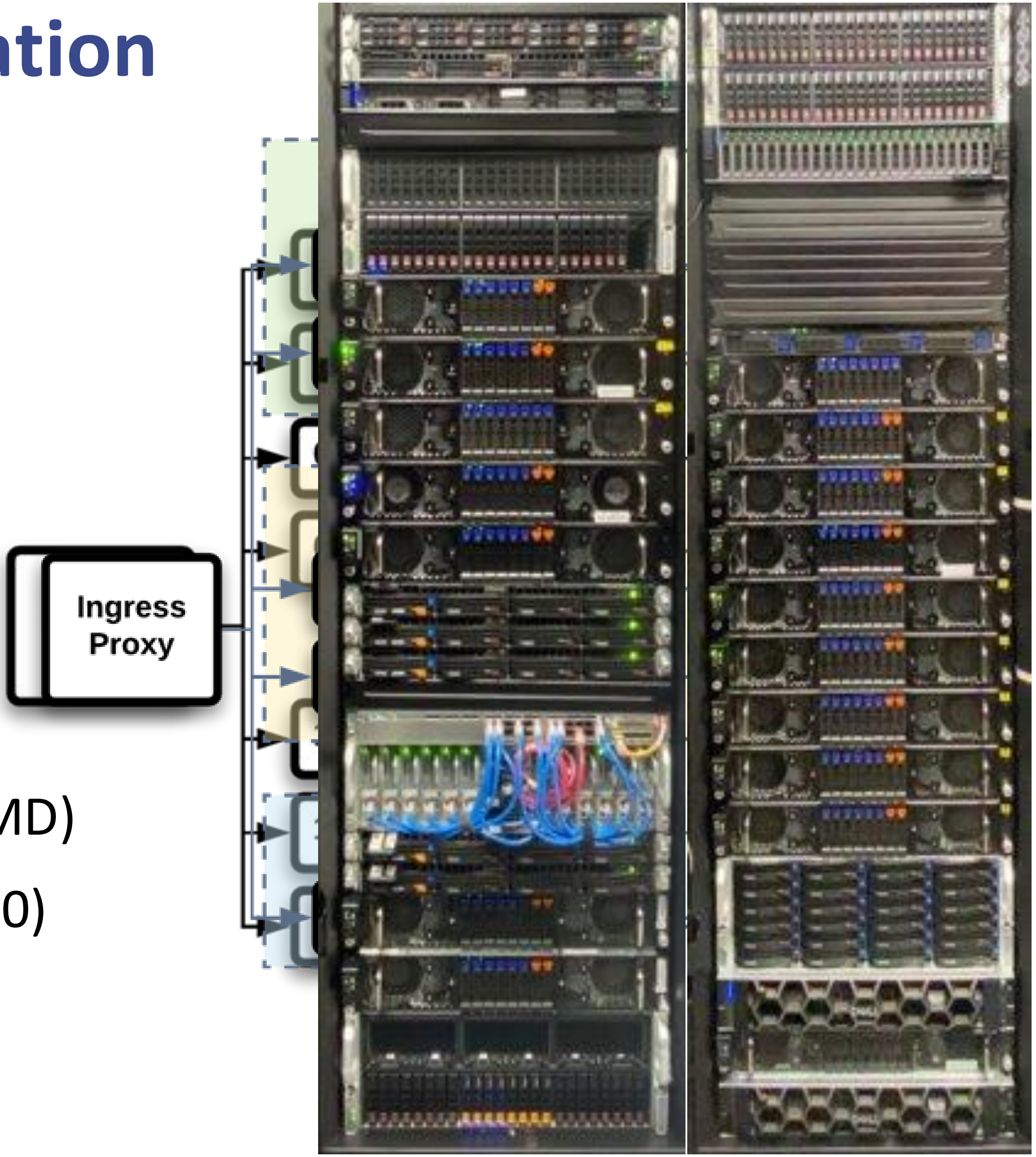
High-level System Organization

Software & OS

- Docker + Kubernetes + Helm
- z2jh + JupyterHub + Jupyter
- Nodes: Debian 12 + Puppet
- Custom rostering/course mgmt

On-Premises Hardware

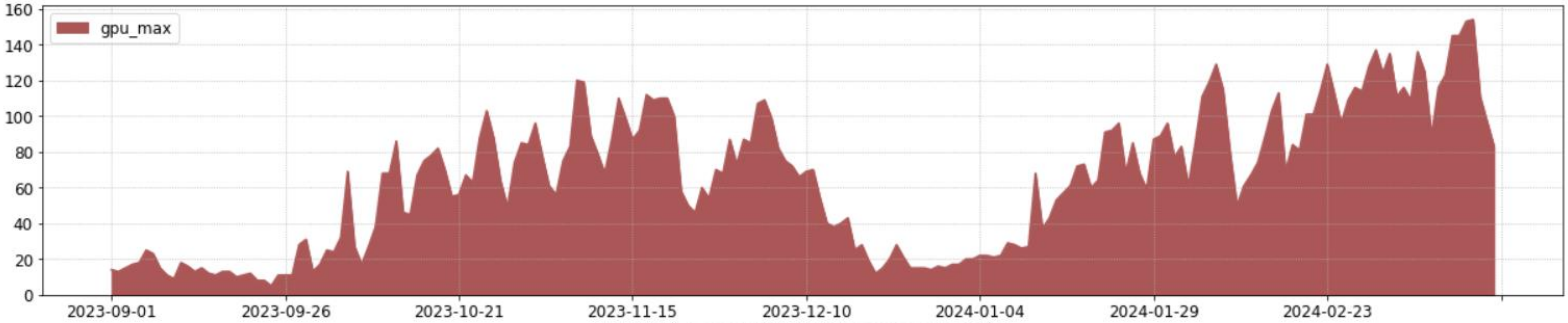
- 23 nodes, 1024 CPU cores (Intel/AMD)
- 132 GPU (1080Ti/2080Ti/A30/A5000)
- 650TB storage via ZFS/NFS
- +& 4 nodes, 24xH100 -> TritonGPT



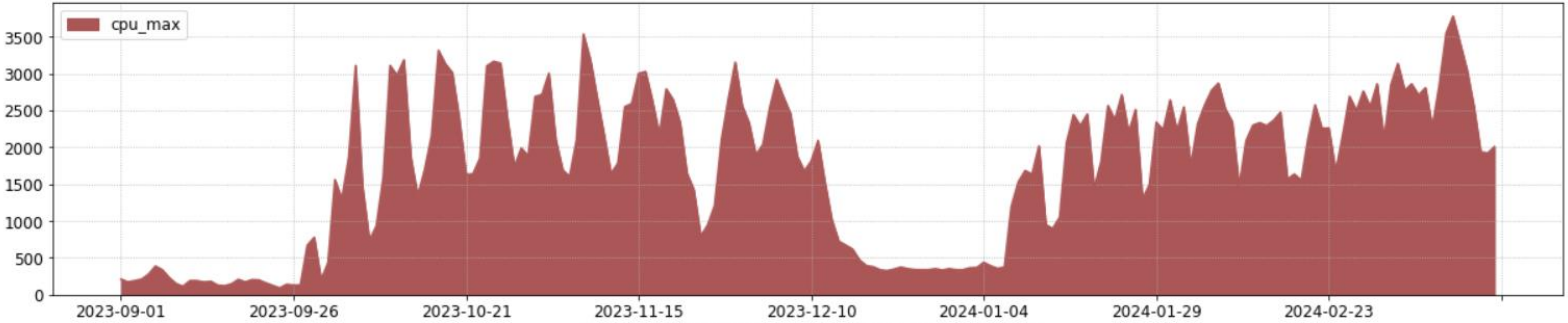
Hardware Outlays, FY2016-17 to Present

		GPU Nodes	CPU Nodes	Network	Filesystem
FY2016-17	$76K	+8 (64*1080Ti)		Arista 10G	add fs01 (80TB flash)
FY2017-18	$60K	+2 (16*1080Ti)			
FY2018-19	$161K	+4 (32*2080Ti)	+3 (64c ea)		
FY2019-20	$61K	+1 (4*RTX Titan)			add fs02 (80TB flash)
FY2020-21	$85K		+3 (128c ea)	Cisco 9k 25G/100G	add fs03 (500TB hdd)
FY2021-22	$83K	+1 (8*A5000)		Cisco mgmt sw	repl fs01 (80TB flash)
FY2022-23	$66K	+1 (4*A30), repl 3*2U			
FY2023-24 (ytd)	$54K	+0 (+12*A30)	*+1 (256c ea)*		
Total ITS	**$645K**				

CPU/GPU Utilization, Sept. 2023 – Mar 2024



DSMLP daily peak GPU; 171793 gpu-hours for period



DSMLP daily peak CPU; 4223995 cpu-core-hours for period

Staffing

- As of 2024, ~4.5 FTE: 0.5 architect, 0.5 manager, 2 x 0.5 sysadmin, 2.5 system engineers (existing staff, no additional funding)
- + team of 5 student staff
- Started out more % architect, fewer sysadmin/engineering staff, no student staff

IT SERVICES

Thank you!

UC San Diego

IT SERVICES

(Additional slides follow if needed)

UC San Diego

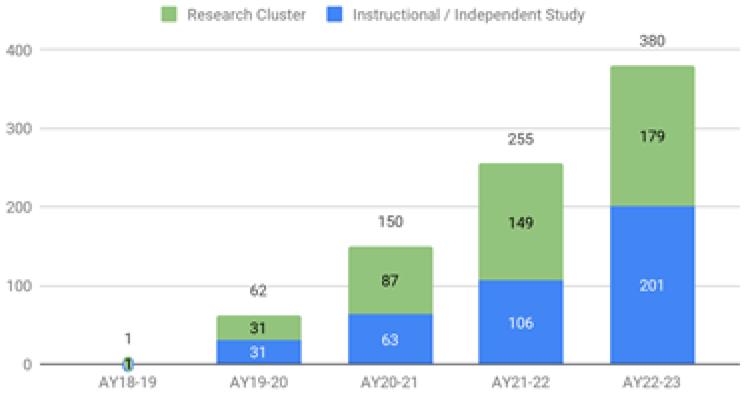
Totals - Instruction

Scheduled Instruction using DSMLP

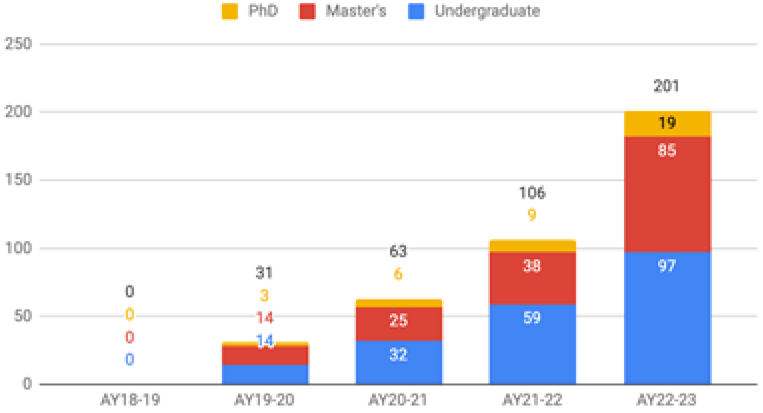
(through Winter 2024)	AY 2023-24	AY 2022-23	Since 2017
Enrollments	10735	16232	71936
Courses	79	138	642
Instructors	56	81	173
Academic Quarters	2	3	24
Divisions/Schools	7	9	9

Research & Independent Study Users

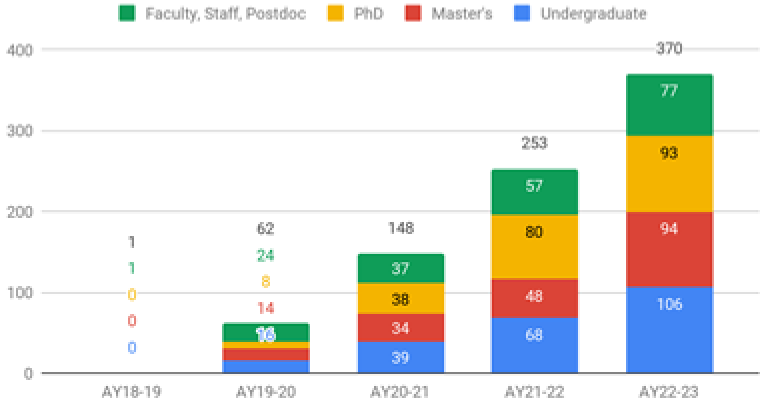
Overall Users (Instructional/ISR & Research Cluster)



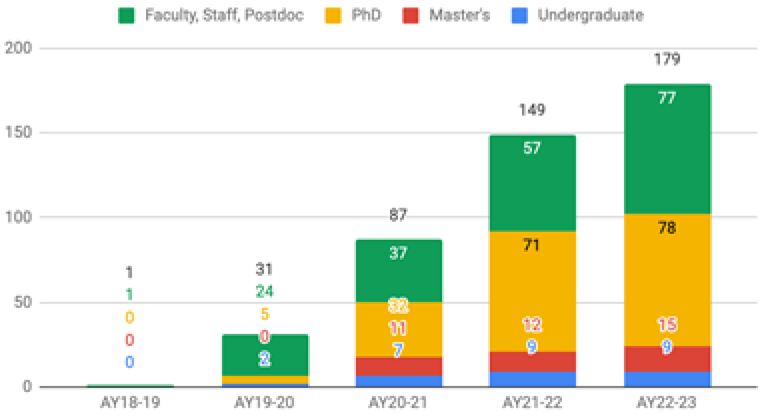
Users by Level (Instructional/ISR)



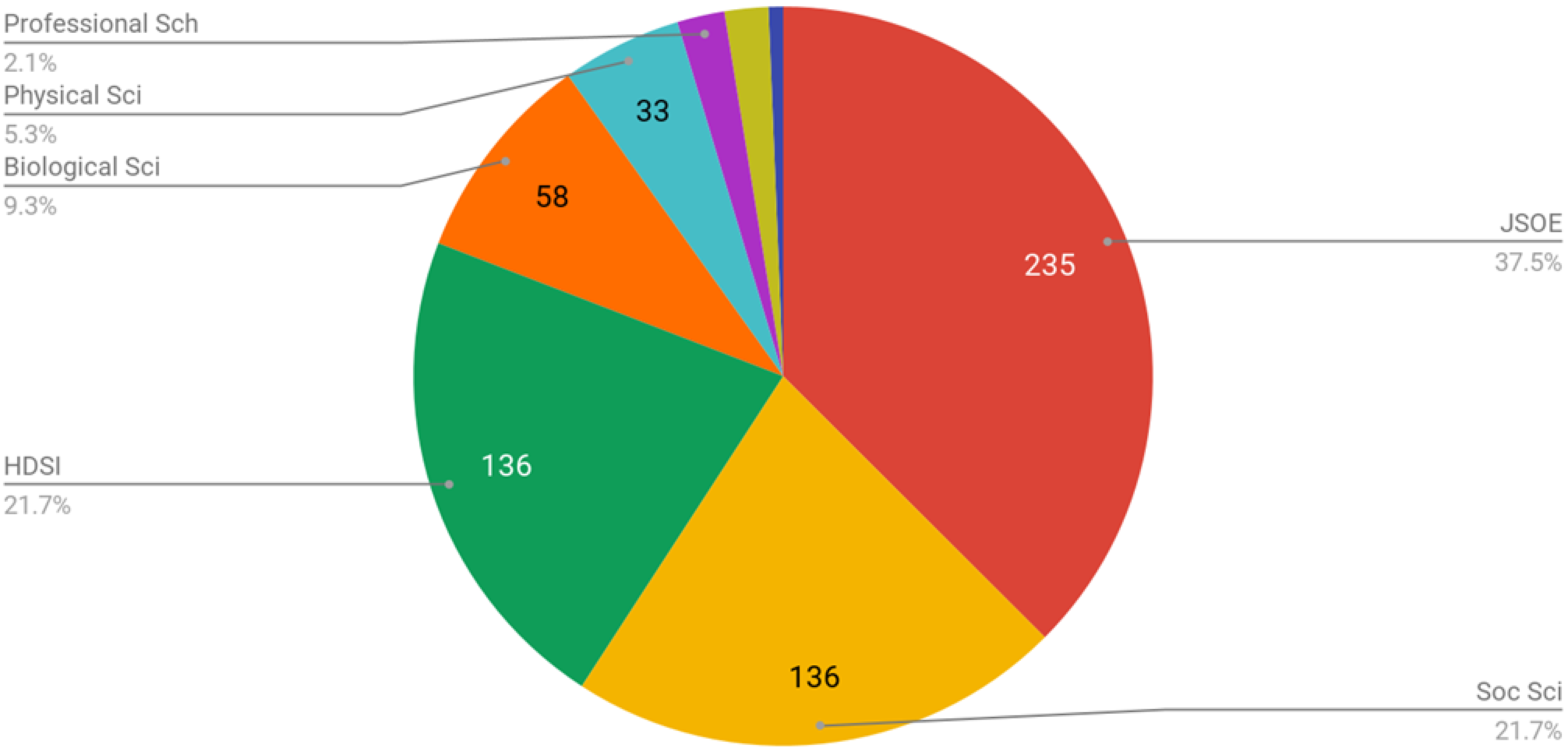
Users by Level (Instructional/ISR & Research Cluster)



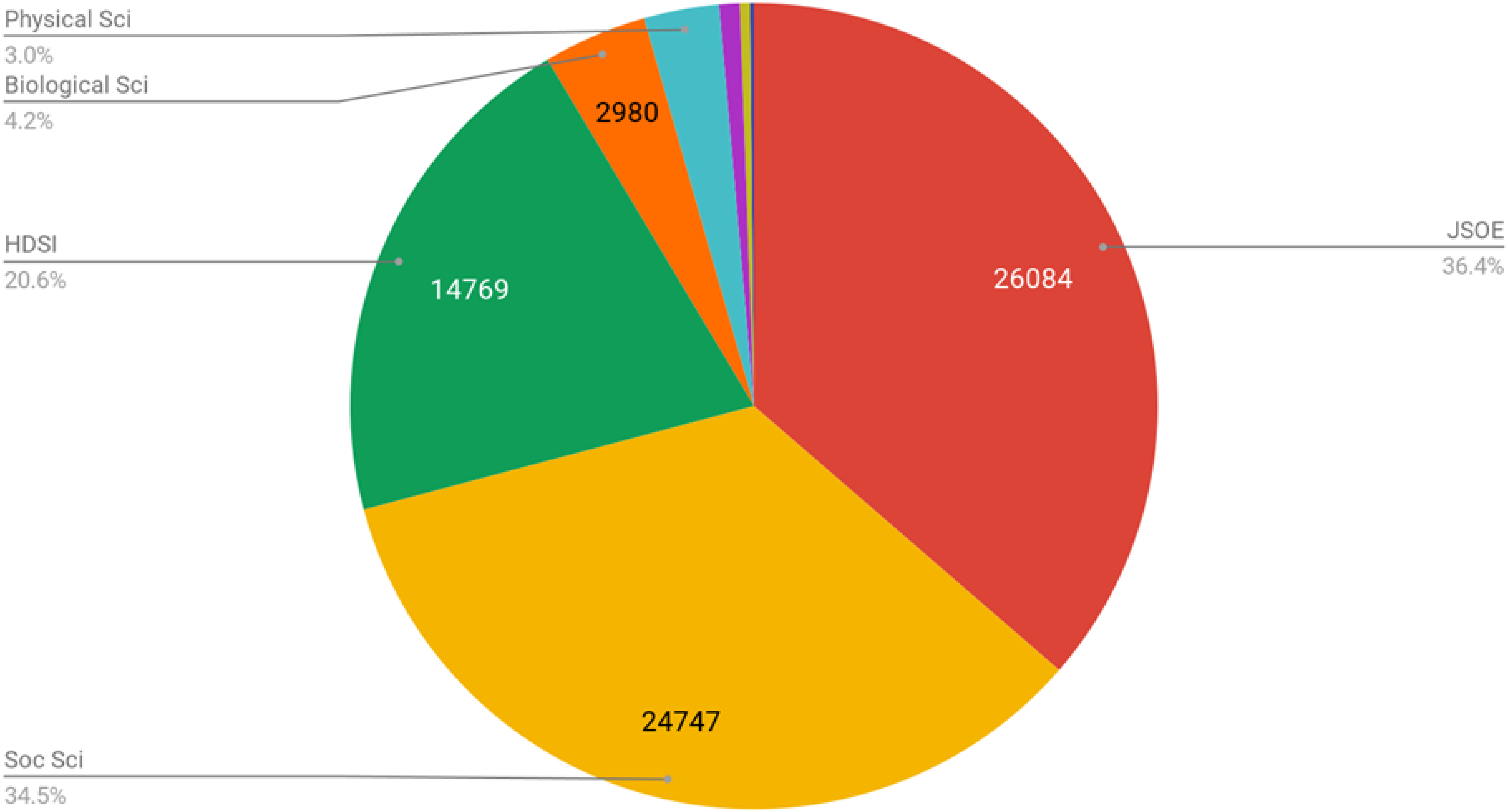
Users by Level (Research Cluster)



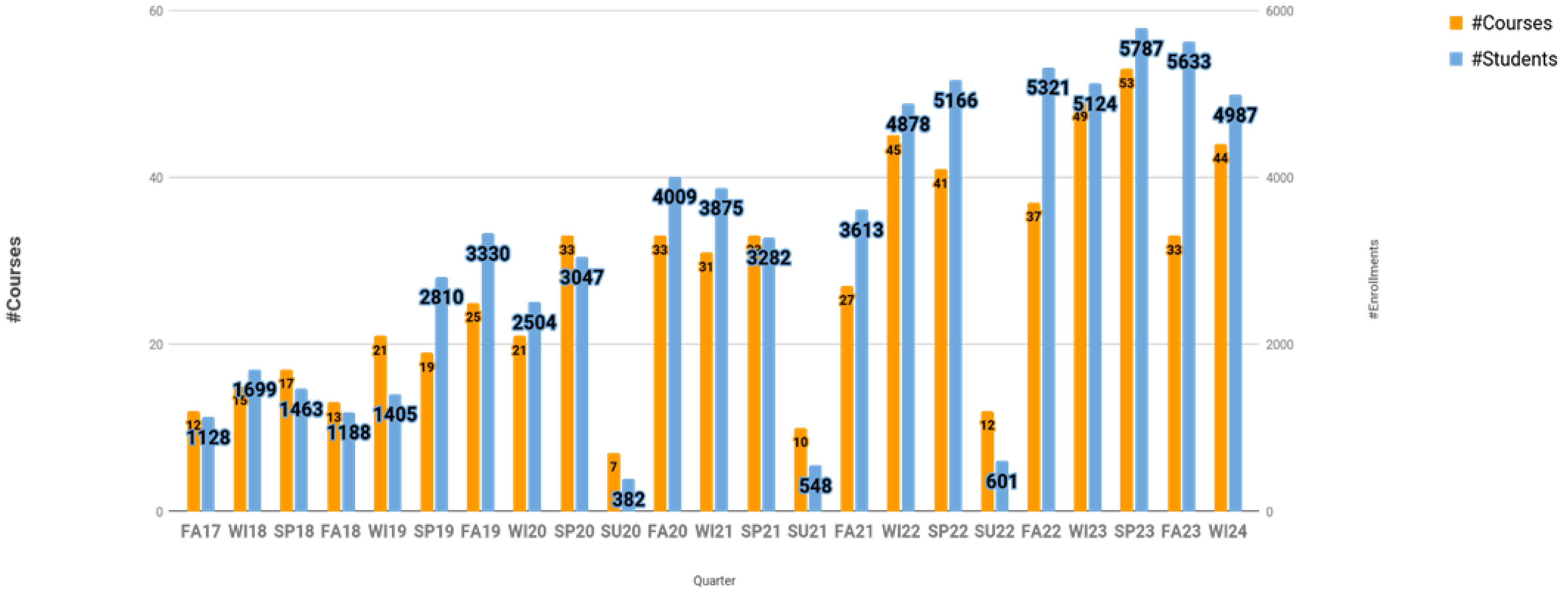
DSMLP Courses by Division, All Time



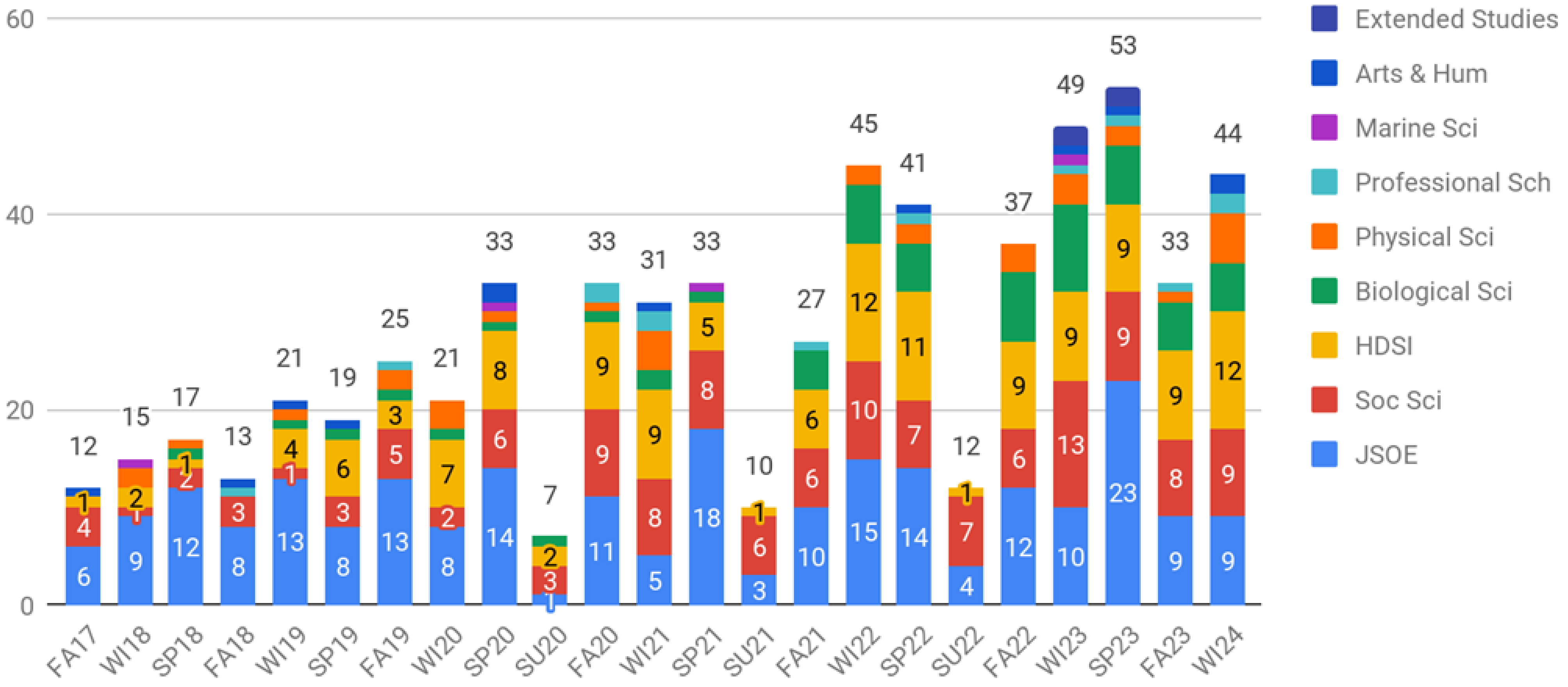
DSMLP Students by Division, All Time



DSMLP Courses, Enrollments by Term

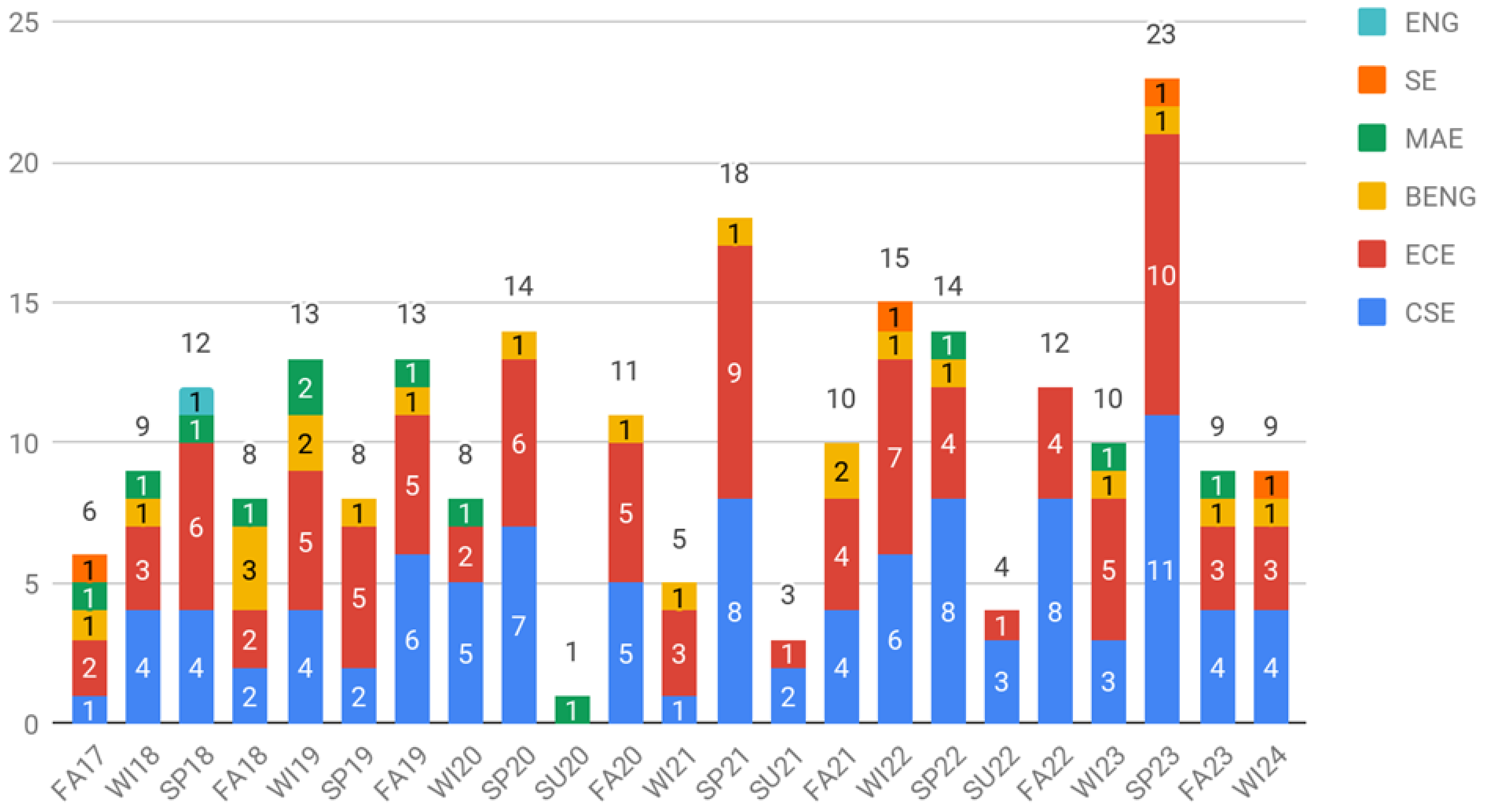


DSMLP Courses by School/Division, Term



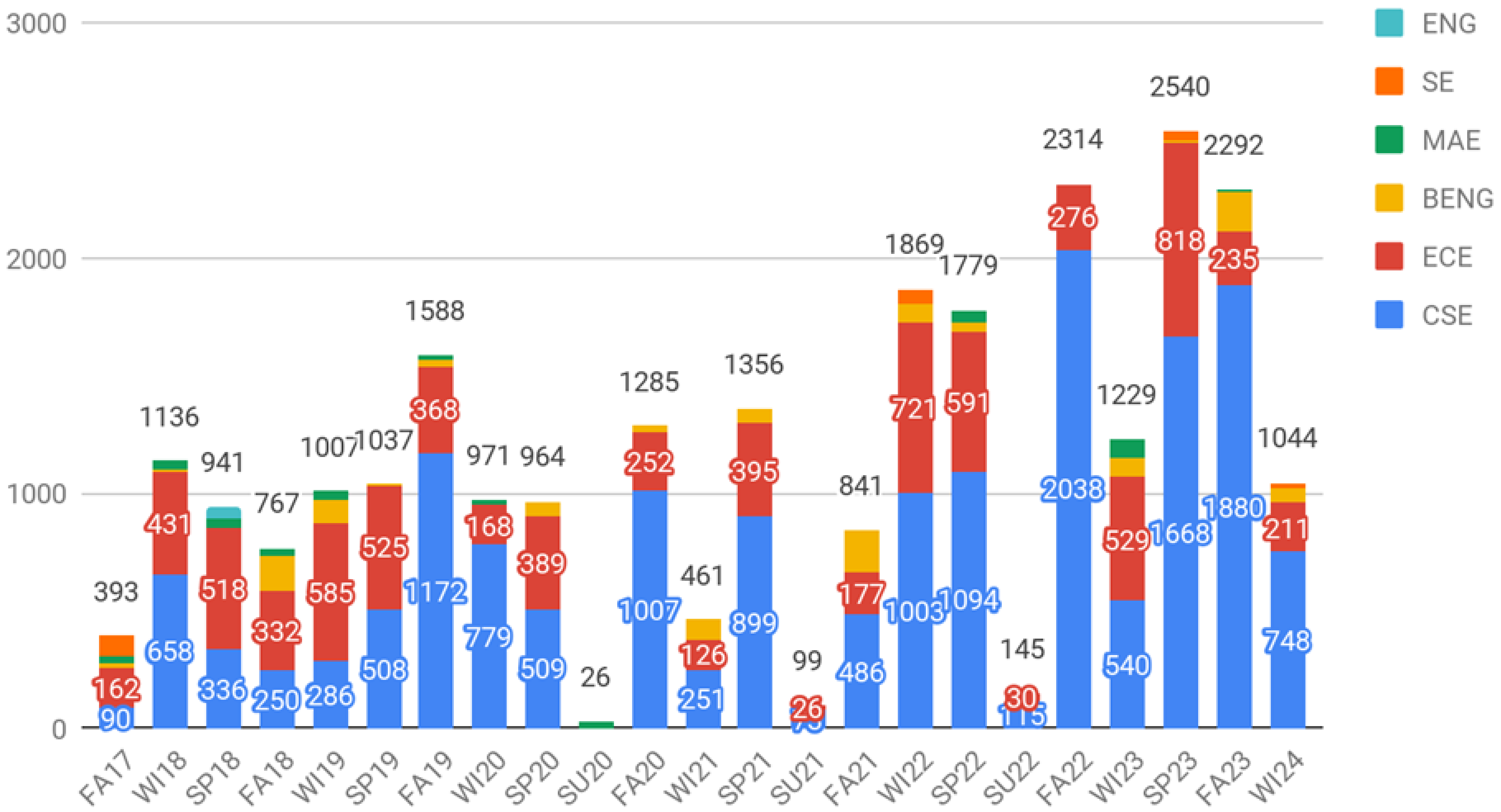
Includes mid-Winter 2024 data

DSMLP Courses (JSOE) by Dept, Term

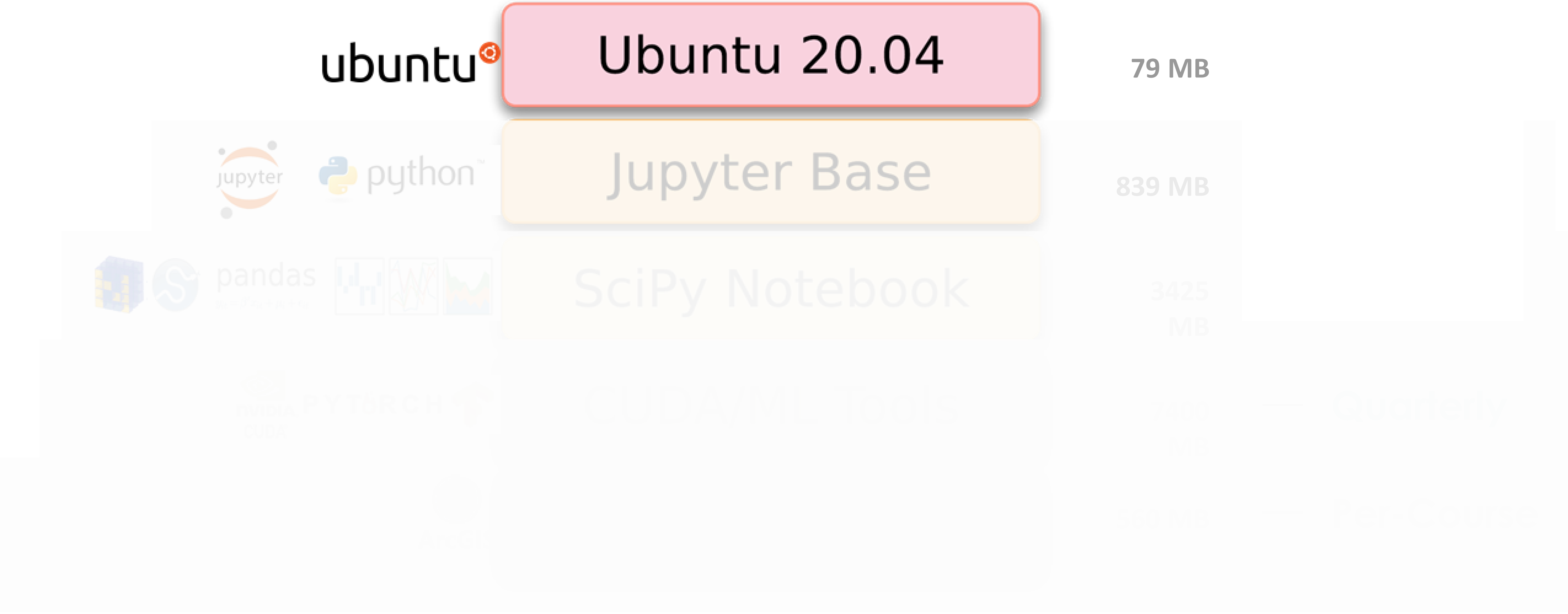


Term	CSE	ECE	BENG	MAE	SE	ENG	Total
FA17	1	2	1	1	1		6
WI18	4	3	1	1			9
SP18	4	6		1		1	12
FA18	2	2	3	1			8
WI19	4	5	2	2			13
SP19	2	5	1				8
FA19	6	5	1	1			13
WI20	5	2		1			8
SP20	7	6	1				14
SU20				1			1
FA20	5	5	1				11
WI21	1	3	1				5
SP21	8	9	1				18
SU21	2	1					3
FA21	4	4	2				10
WI22	6	7	1		1		15
SP22	8	4	1	1			14
SU22	3	1					4
FA22	8	4					12
WI23	3	5	1	1			10
SP23	11	10	1		1		23
FA23	4	3	1	1			9
WI24	4	3	1		1		9

DSMLP Students (JSOE) by Dept, Term



Course-Custom Software Stacks via Docker



Ubuntu 20.04

79 MB

Jupyter Base

839 MB

SciPy Notebook

3425 MB

CUDA/ML Tools

7400 MB

Quarterly

560 MB

Per-Course