

# **EE3801 Data Engineering Principles**

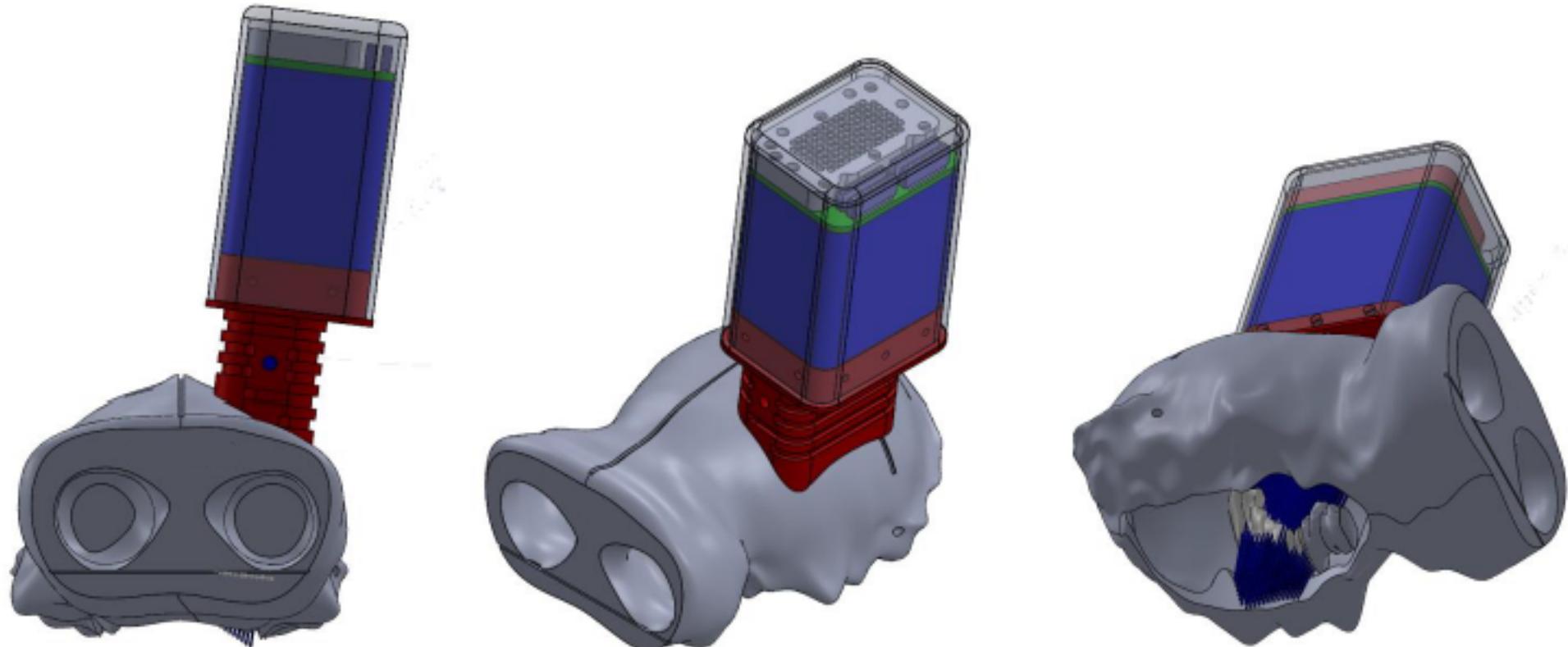
**Shih-Cheng YEN**

# Topics

- Vectorization techniques in data processing and analysis
- Data visualization
- Amazon Web Services Elastic Compute Cloud
- Elastic Block System file system
- Parallel Cluster
- SLURM Workload Manager
- Data pipeline construction, optimization, and operation
- High-throughput data visualization and inspection

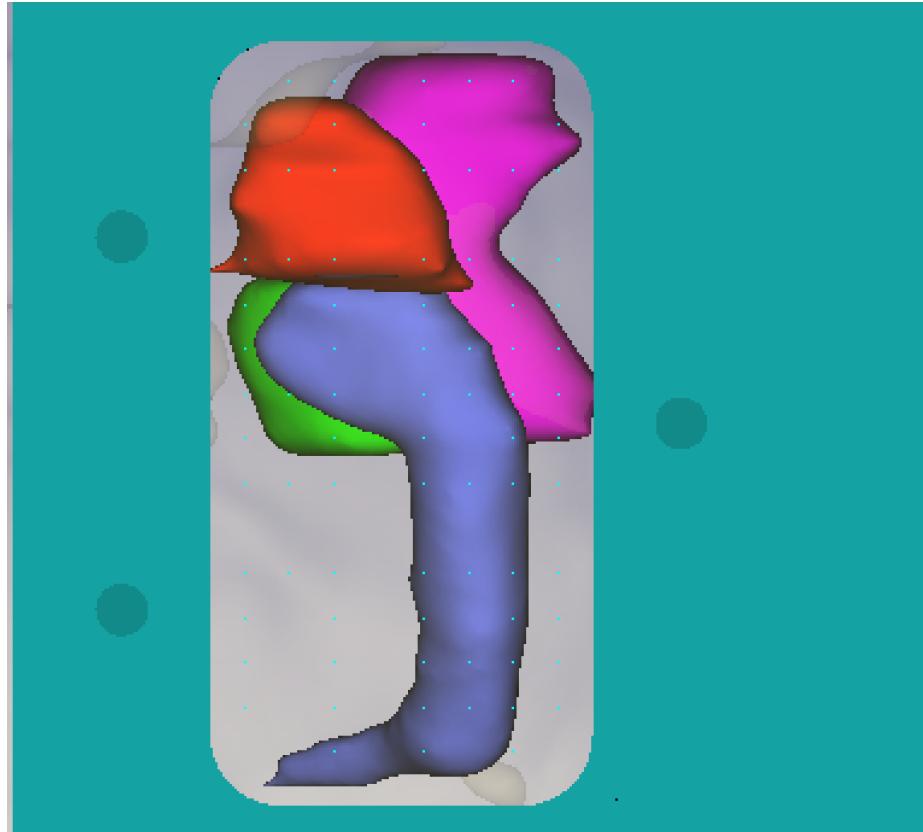
# Dataset

Intra-cranial neural recordings with > 100 electrodes



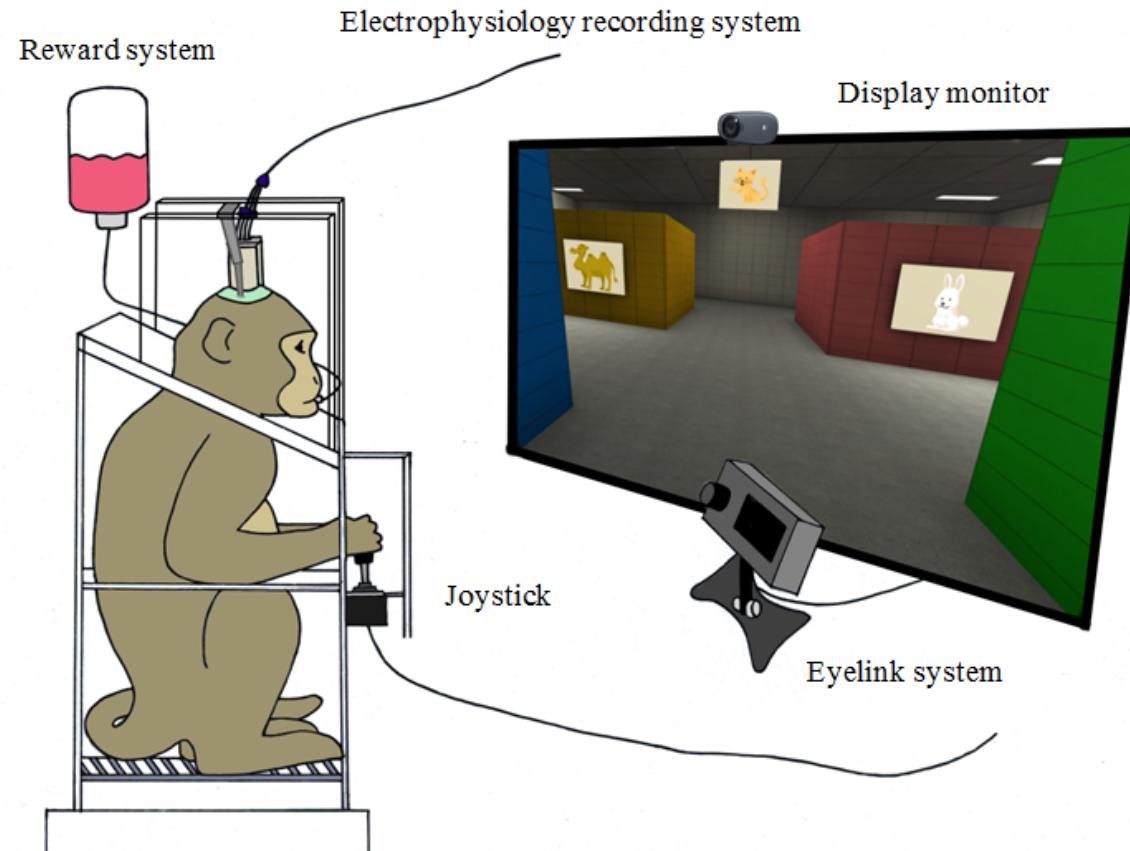
# Dataset

**Intra-cranial neural recordings with > 100 electrodes**



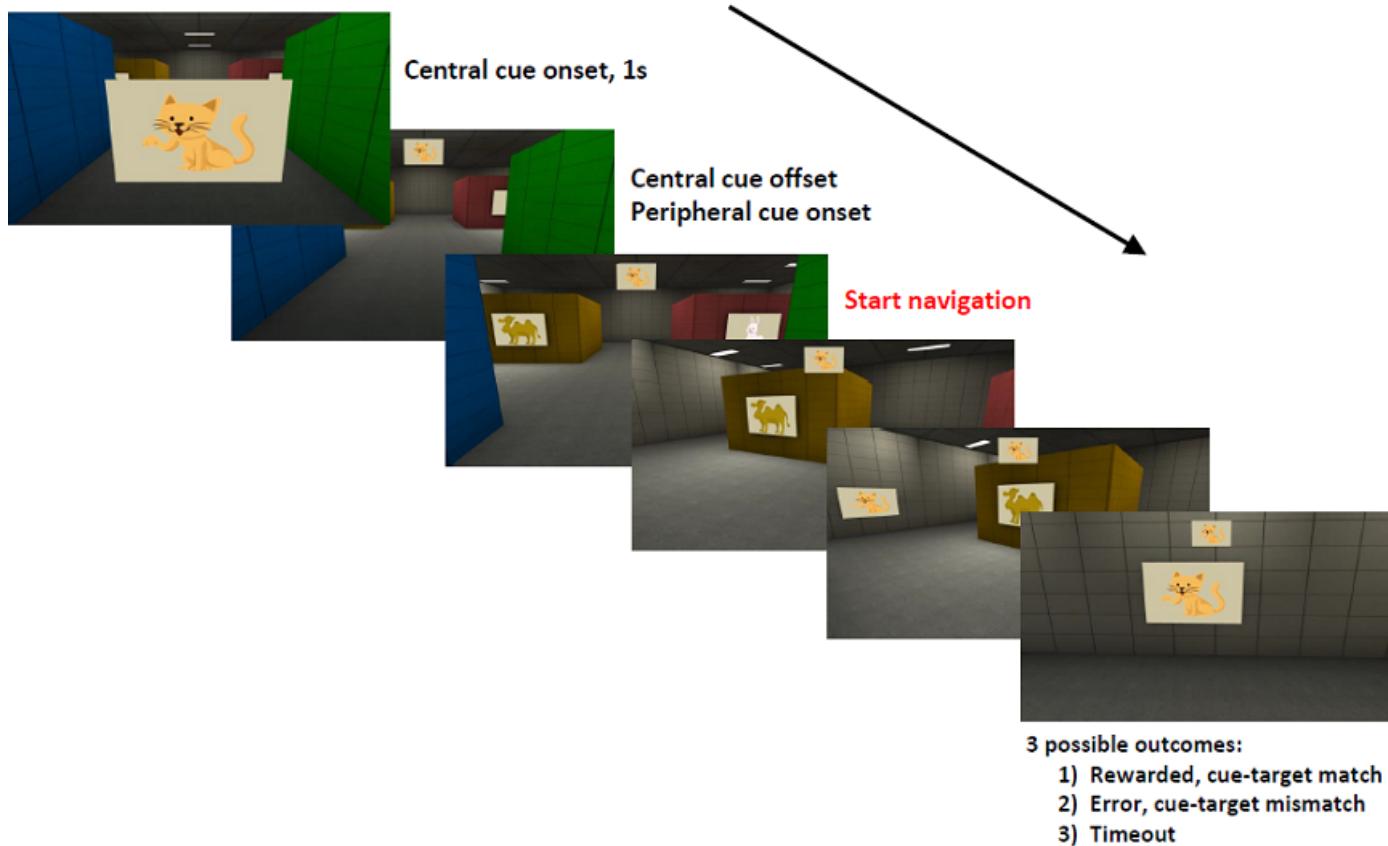
# Dataset

## Awake behaving animal



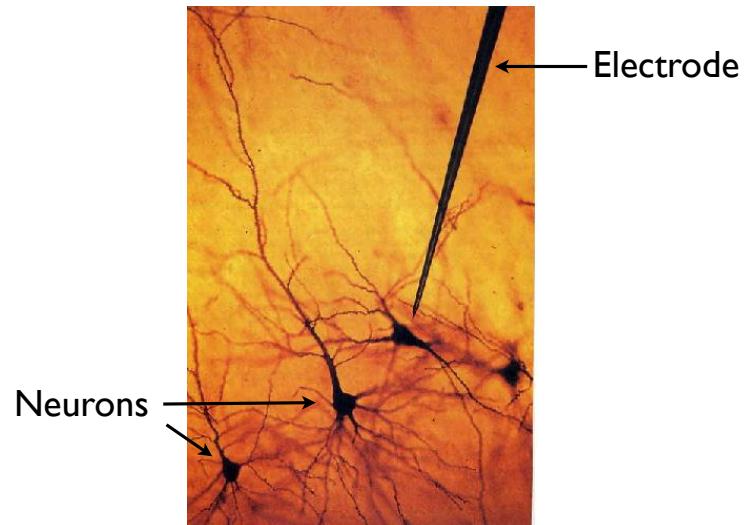
# Dataset

## Experimental task

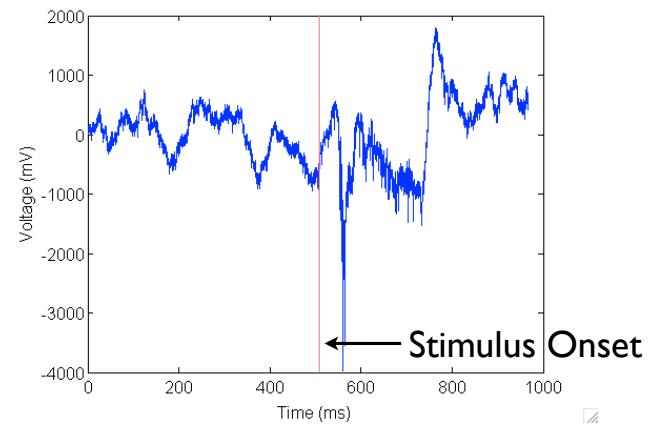


# Dataset

## Neural recordings



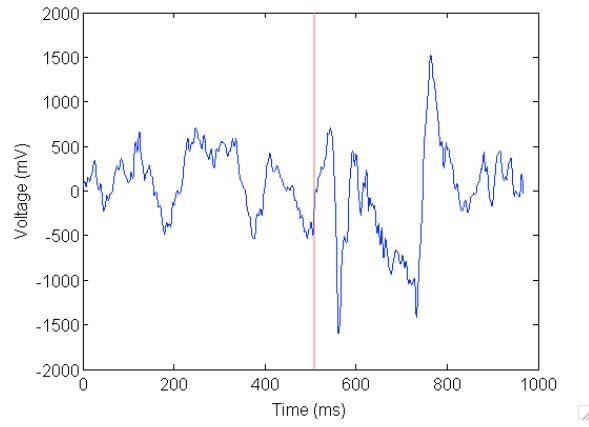
Capable of recording from multiple neurons



Raw signal is amplified 4000 times and contains frequencies up to 10 kHz, so data is recorded at 30 kHz. The data file for each day is  $\approx$  40 GB.

# Dataset

## Neural recordings

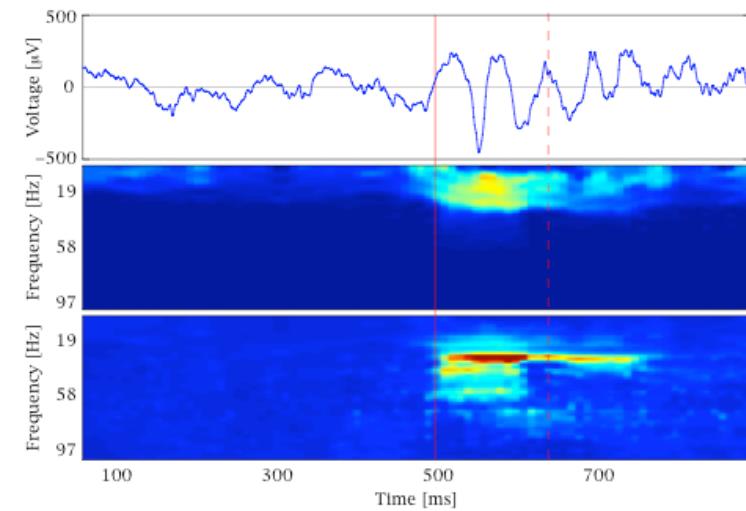


Low-pass filtered at  $f = 150$  Hz  
Local Field Potential - population synaptic activity

LFP

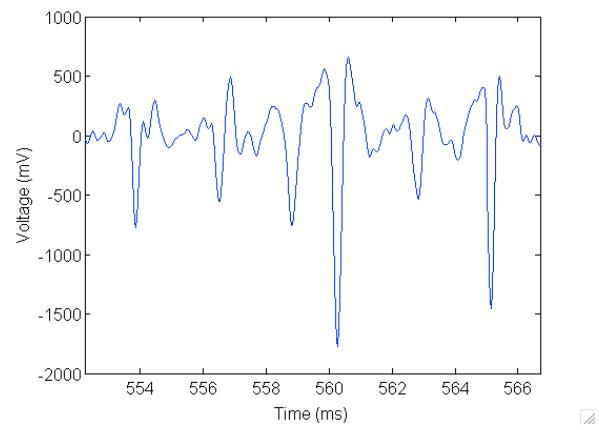
Spectrogram

Normalized Spectrogram

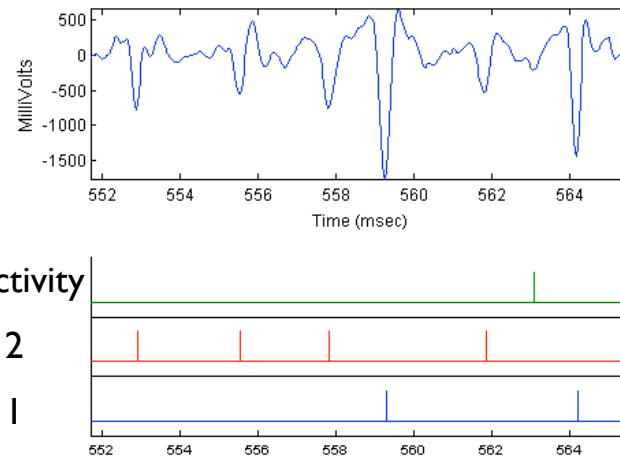


# Dataset

## Neural recordings

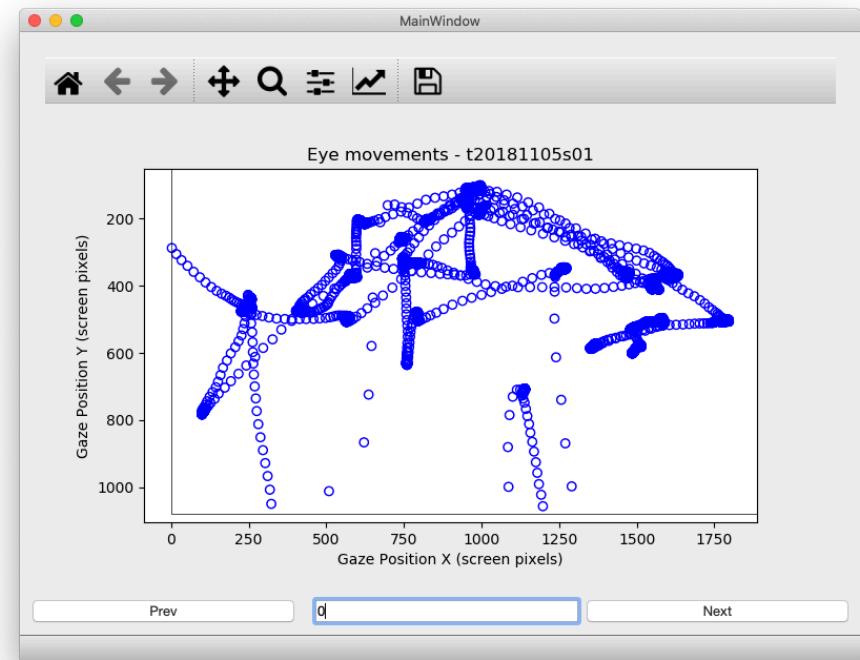
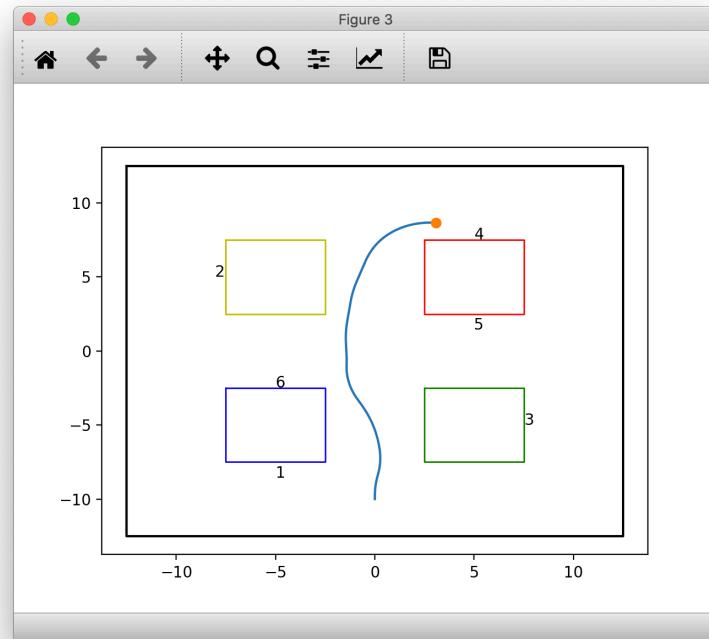


Spikes from multiple neurons present



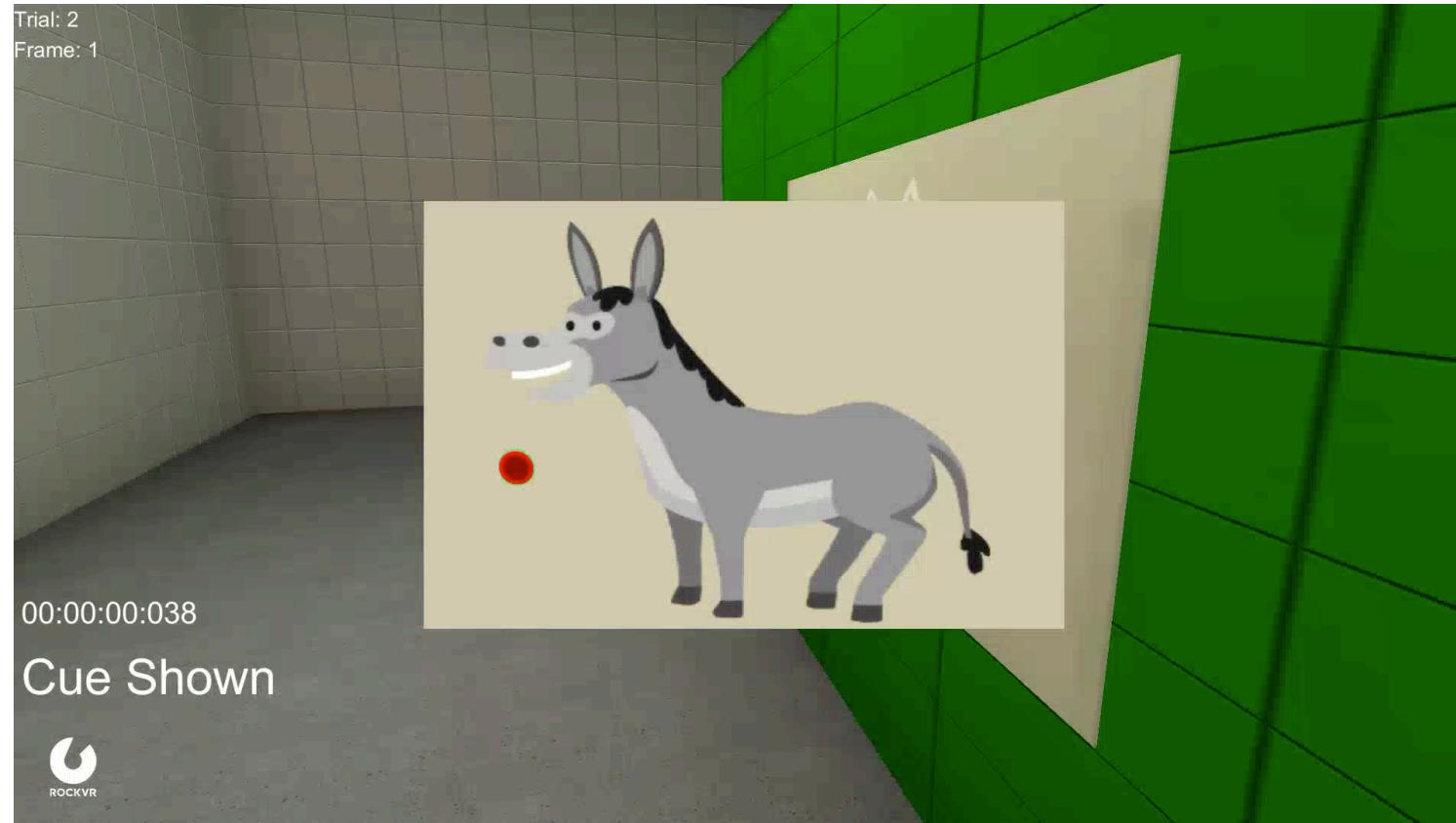
# Dataset

## Position in Unity3D maze & eye-position

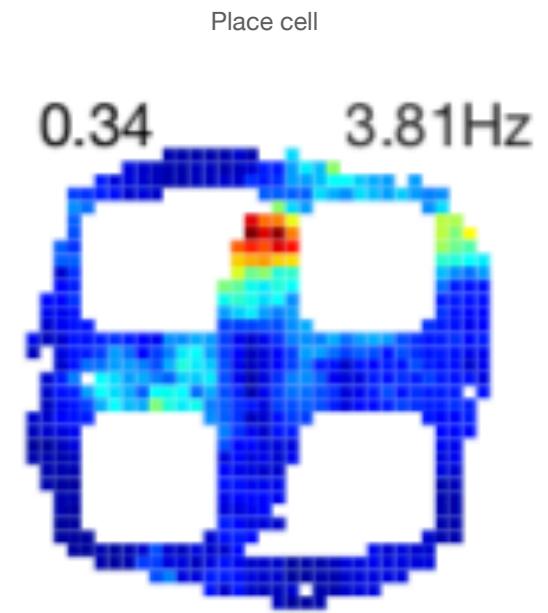
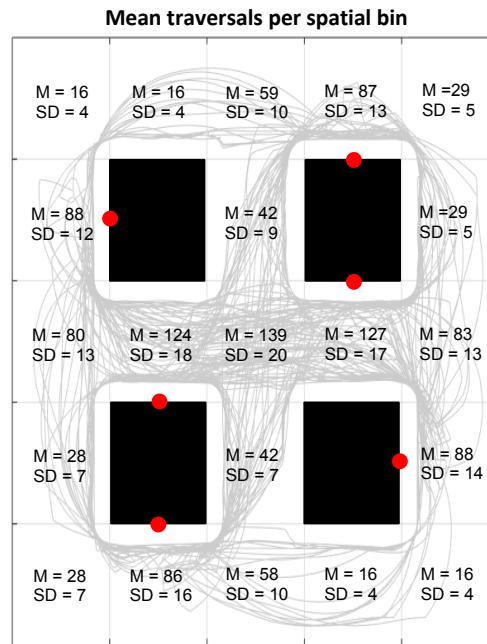
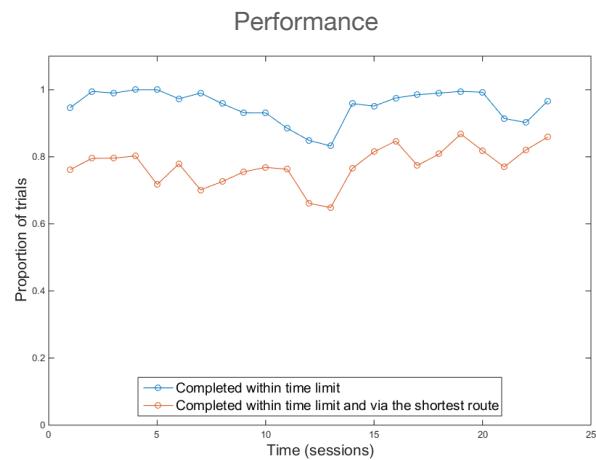


# Dataset

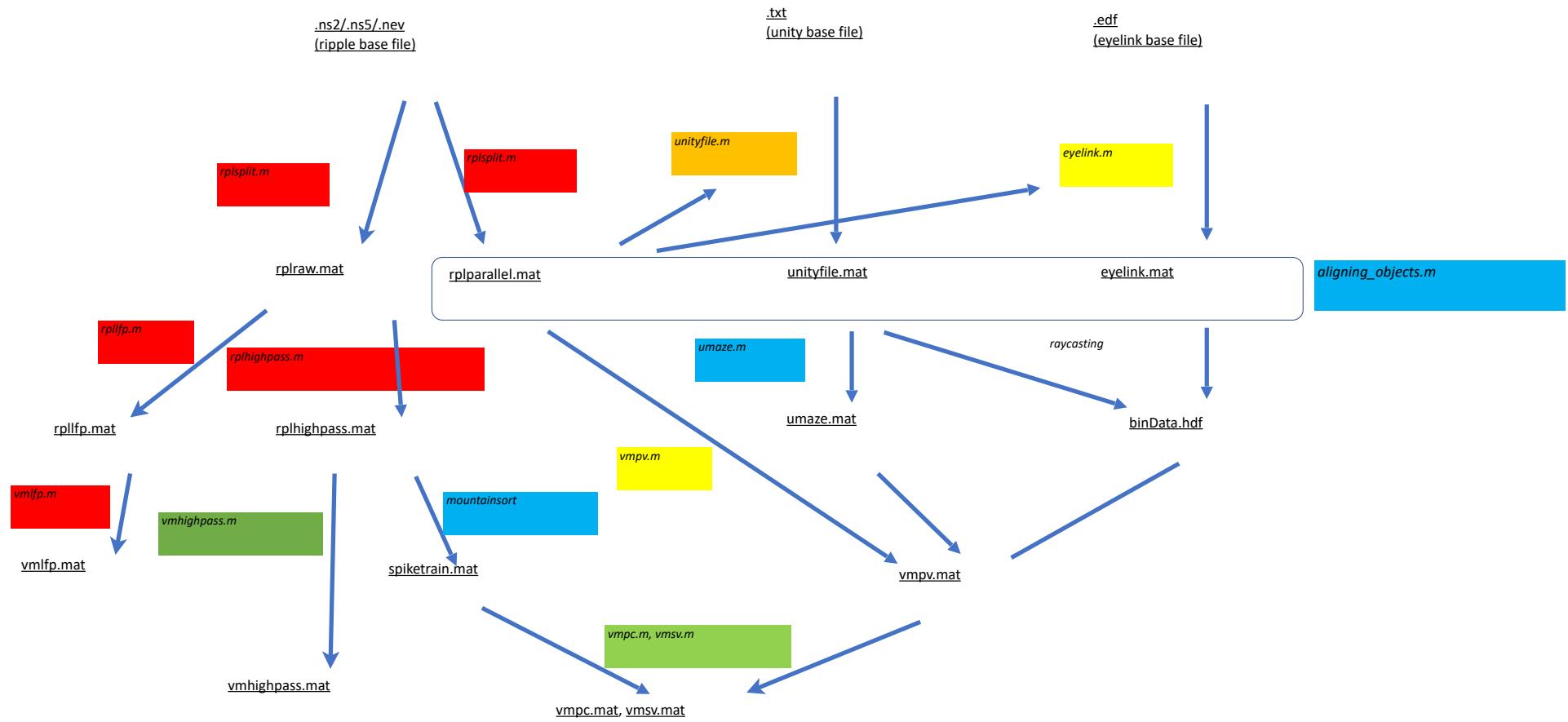
## Raycasting



# Dataset Results



# Data pipeline



# Parallel Data Processing

## Parallel Processing

Serial Pipeline

Dependencies	RPLParallel	RPLSplit	RPLLFP	RPLHighPass	Unity	EDFSplit
RPLParallel						
RPLSplit						
RPLLFP		✓				
RPLHighPass		✓				
Unity	✓					
EDFSplit	✓					
aligning_objects	✓			✓	✓	
raycast				✓	✓	
mountain_batch			✓			

Parallel Pipeline

RPLParallel	RPLSplit
Unity	RPLLFP
EDFSplit	RPLHighPass
aligning_objects	mountain_batch
raycast	

# Optimizing Parallel Processing

## Parallel Processing (2 jobs)

RPLParallel	RPLSplit		
Unity	session01	channel001	rplraw_xxxx.hkl
EDFSplit	sessioneye	channel002	rplraw_xxxx.hkl
aligning_objects		...	...
raycast		channel124	rplraw_xxxx.hkl
	RPLLFP		
	session01	channel001	rpllfp_xxxx.hkl
	sessioneye	channel002	rpllfp_xxxx.hkl
		...	...
		channel124	rpllfp_xxxx.hkl
	RPLHighPass		
	session01	channel001	rplhighpass_xxxx.hkl
	sessioneye	channel002	rplhighpass_xxxx.hkl
		...	...
		channel124	rplhighpass_xxxx.hkl
	mountain_batch		
	session01	channel001	firings.mda
		channel002	firings.mda
		...	...
		channel124	firings.mda

Time ↓

# Optimizing Parallel Processing

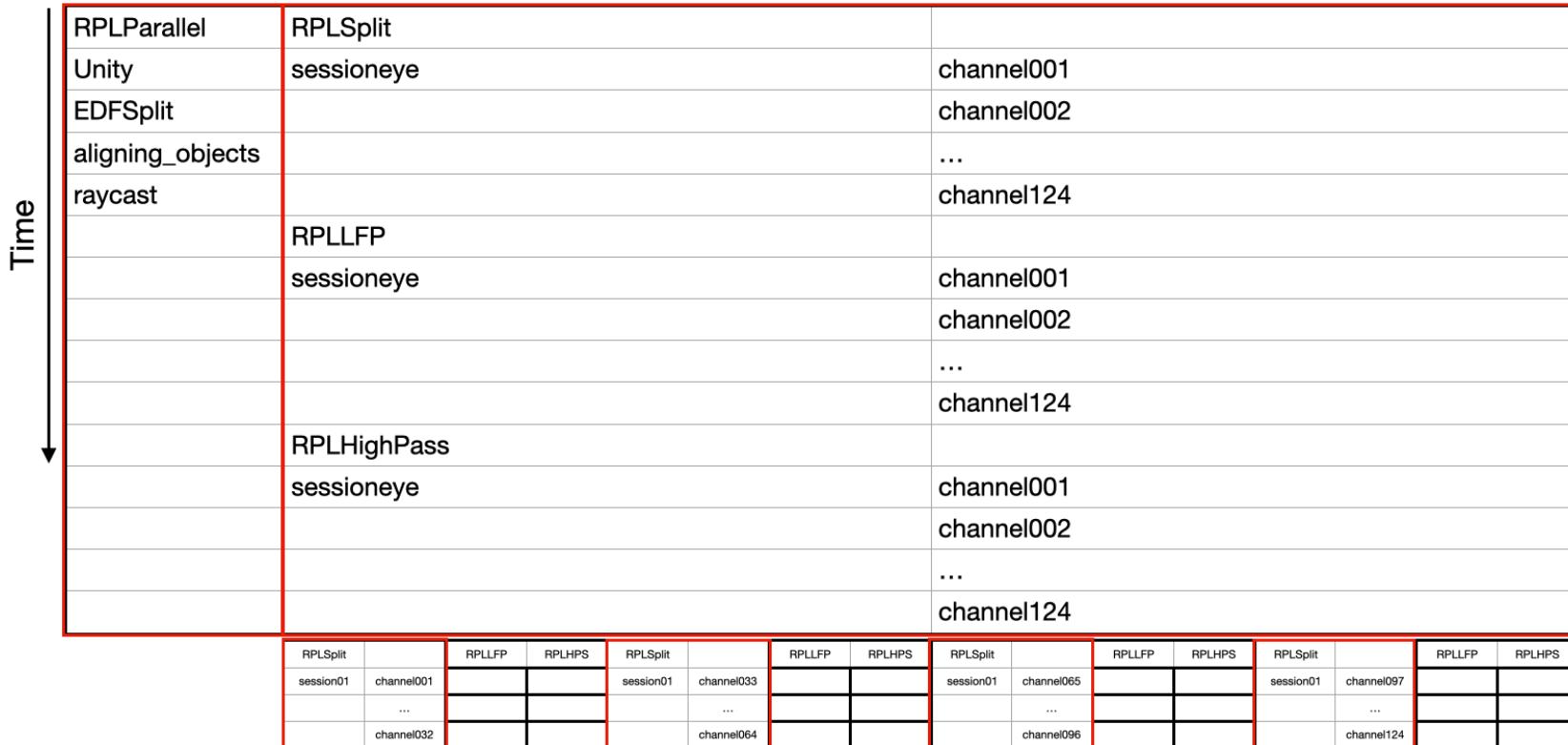
## Coarse-Grained Parallel Processing (5 jobs)

Time ↓

RPLParallel	RPLSplit		RPLSplit		RPLSplit		RPLSplit	
Unity	session01	channel001	session01	channel033	session01	channel065	session01	channel097
EDFSplit	sessioneye	channel002	sessioneye	channel034	sessioneye	channel066	sessioneye	channel098
<td>...</td> <td></td> <td>...</td> <td></td> <td>...</td> <td></td> <td>...</td> <td></td>	...		...		...		...	
raycast		channel032		channel064		channel096		channel124
	RPLLFP		RPLLFP		RPLLFP		RPLLFP	
	session01	channel001	session01	channel033	session01	channel065	session01	channel097
	sessioneye	channel002	sessioneye	channel034	sessioneye	channel066	sessioneye	channel098
	...		...		...		...	
		channel032		channel064		channel096		channel124
	RPLHighPass		RPLHighPass		RPLHighPass		RPLHighPass	
	session01	channel001	session01	channel033	session01	channel065	session01	channel097
	sessioneye	channel002	sessioneye	channel034	sessioneye	channel066	sessioneye	channel098
	...		...		...		...	
		channel032		channel064		channel096		channel124
	mountain_batch		mountain_batch		mountain_batch		mountain_batch	
	session01	channel001	session01	channel033	session01	channel065	session01	channel097
		channel002		channel034		channel066		channel098
	...		...		...		...	
		channel032		channel064		channel096		channel124

# Optimizing Parallel Processing

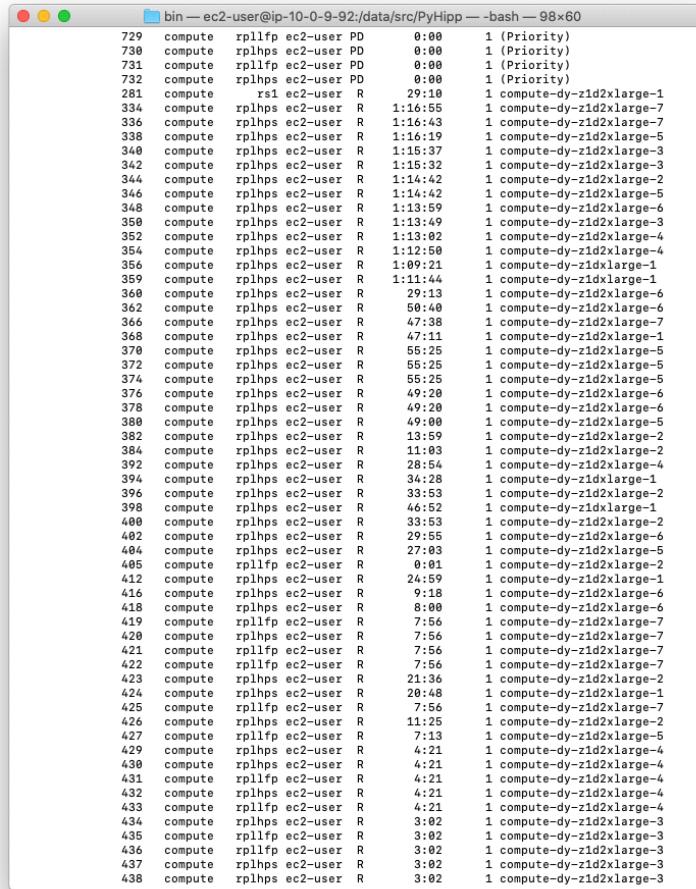
## Fine-Grained Parallel Processing



Jobs:  $6 + 110 + 110 = 226$

# Optimizing Parallel Processing

## Fine-Grained Parallel Processing (226 jobs)



The screenshot shows a terminal window titled "bin — ec2-user@ip-10-0-9-92:/data/src/PyHipp — bash — 98x60". The window displays a list of 226 compute tasks. Each task is identified by a job ID (e.g., 729, 730, ..., 438), followed by the command "compute", the function name ("rpl1fp" or "rpl1fps"), the user ("ec2-user"), the path ("PD" or "R"), and the start time (e.g., 0:00, 1:16:55, 1:16:43). To the right of each task, there is a column of numbers representing priority levels, ranging from 1 (Priority) at the top to 7 at the bottom.

Job ID	Command	User	Path	Start Time	Priority	
729	compute	rpl1fp	ec2-user	PD	0:00	1 (Priority)
730	compute	rpl1fps	ec2-user	PD	0:00	1 (Priority)
731	compute	rpl1fp	ec2-user	PD	0:00	1 (Priority)
732	compute	rpl1fps	ec2-user	PD	0:00	1 (Priority)
281	compute	rs1	ec2-user	R	29:10	1 compute-dy-z1d2xlarge-1
334	compute	rpl1fps	ec2-user	R	1:16:55	1 compute-dy-z1d2xlarge-7
336	compute	rpl1fps	ec2-user	R	1:16:43	1 compute-dy-z1d2xlarge-7
338	compute	rpl1fps	ec2-user	R	1:16:19	1 compute-dy-z1d2xlarge-5
340	compute	rpl1fps	ec2-user	R	1:15:37	1 compute-dy-z1d2xlarge-3
342	compute	rpl1fps	ec2-user	R	1:15:32	1 compute-dy-z1d2xlarge-3
344	compute	rpl1fps	ec2-user	R	1:14:42	1 compute-dy-z1d2xlarge-2
346	compute	rpl1fps	ec2-user	R	1:14:42	1 compute-dy-z1d2xlarge-5
348	compute	rpl1fps	ec2-user	R	1:13:59	1 compute-dy-z1d2xlarge-6
350	compute	rpl1fps	ec2-user	R	1:13:49	1 compute-dy-z1d2xlarge-3
352	compute	rpl1fps	ec2-user	R	1:13:02	1 compute-dy-z1d2xlarge-4
354	compute	rpl1fps	ec2-user	R	1:12:50	1 compute-dy-z1d2xlarge-4
356	compute	rpl1fps	ec2-user	R	1:09:21	1 compute-dy-z1d2xlarge-1
359	compute	rpl1fps	ec2-user	R	1:11:44	1 compute-dy-z1d2xlarge-1
360	compute	rpl1fps	ec2-user	R	29:13	1 compute-dy-z1d2xlarge-6
362	compute	rpl1fps	ec2-user	R	50:48	1 compute-dy-z1d2xlarge-6
366	compute	rpl1fps	ec2-user	R	47:38	1 compute-dy-z1d2xlarge-7
368	compute	rpl1fps	ec2-user	R	47:11	1 compute-dy-z1d2xlarge-1
370	compute	rpl1fps	ec2-user	R	55:25	1 compute-dy-z1d2xlarge-5
372	compute	rpl1fps	ec2-user	R	55:25	1 compute-dy-z1d2xlarge-5
374	compute	rpl1fps	ec2-user	R	55:25	1 compute-dy-z1d2xlarge-5
376	compute	rpl1fps	ec2-user	R	49:20	1 compute-dy-z1d2xlarge-6
378	compute	rpl1fps	ec2-user	R	49:20	1 compute-dy-z1d2xlarge-6
380	compute	rpl1fps	ec2-user	R	49:00	1 compute-dy-z1d2xlarge-5
382	compute	rpl1fps	ec2-user	R	13:59	1 compute-dy-z1d2xlarge-2
384	compute	rpl1fps	ec2-user	R	11:03	1 compute-dy-z1d2xlarge-2
392	compute	rpl1fps	ec2-user	R	28:54	1 compute-dy-z1d2xlarge-4
394	compute	rpl1fps	ec2-user	R	34:28	1 compute-dy-z1d2xlarge-1
396	compute	rpl1fps	ec2-user	R	33:53	1 compute-dy-z1d2xlarge-2
398	compute	rpl1fps	ec2-user	R	46:52	1 compute-dy-z1d2xlarge-1
400	compute	rpl1fps	ec2-user	R	33:53	1 compute-dy-z1d2xlarge-2
402	compute	rpl1fps	ec2-user	R	29:55	1 compute-dy-z1d2xlarge-6
404	compute	rpl1fps	ec2-user	R	27:03	1 compute-dy-z1d2xlarge-5
405	compute	rpl1fp	ec2-user	R	0:01	1 compute-dy-z1d2xlarge-2
412	compute	rpl1fps	ec2-user	R	24:59	1 compute-dy-z1d2xlarge-1
416	compute	rpl1fps	ec2-user	R	9:18	1 compute-dy-z1d2xlarge-6
418	compute	rpl1fps	ec2-user	R	8:00	1 compute-dy-z1d2xlarge-6
419	compute	rpl1fp	ec2-user	R	7:56	1 compute-dy-z1d2xlarge-7
420	compute	rpl1fps	ec2-user	R	7:56	1 compute-dy-z1d2xlarge-7
421	compute	rpl1fp	ec2-user	R	7:56	1 compute-dy-z1d2xlarge-7
422	compute	rpl1fp	ec2-user	R	7:56	1 compute-dy-z1d2xlarge-7
423	compute	rpl1fps	ec2-user	R	21:36	1 compute-dy-z1d2xlarge-2
424	compute	rpl1fps	ec2-user	R	20:48	1 compute-dy-z1d2xlarge-1
425	compute	rpl1fp	ec2-user	R	7:56	1 compute-dy-z1d2xlarge-7
426	compute	rpl1fps	ec2-user	R	11:25	1 compute-dy-z1d2xlarge-2
427	compute	rpl1fp	ec2-user	R	7:13	1 compute-dy-z1d2xlarge-5
429	compute	rpl1fps	ec2-user	R	4:21	1 compute-dy-z1d2xlarge-4
430	compute	rpl1fps	ec2-user	R	4:21	1 compute-dy-z1d2xlarge-4
431	compute	rpl1fp	ec2-user	R	4:21	1 compute-dy-z1d2xlarge-4
432	compute	rpl1fps	ec2-user	R	4:21	1 compute-dy-z1d2xlarge-4
433	compute	rpl1fp	ec2-user	R	4:21	1 compute-dy-z1d2xlarge-4
434	compute	rpl1fps	ec2-user	R	3:02	1 compute-dy-z1d2xlarge-3
435	compute	rpl1fp	ec2-user	R	3:02	1 compute-dy-z1d2xlarge-3
436	compute	rpl1fp	ec2-user	R	3:02	1 compute-dy-z1d2xlarge-3
437	compute	rpl1fps	ec2-user	R	3:02	1 compute-dy-z1d2xlarge-3
438	compute	rpl1fps	ec2-user	R	3:02	1 compute-dy-z1d2xlarge-3

Cannot actually run 64 jobs as the Master Node uses 2 CPUs and the EC2 instance also uses 2 CPUs

## Lesson Plan – Part 2

Week	Monday	Thursday
7	Quiz 1 (Part 1)	Lecture
8	Lab 4	Lecture
9	Lab 5	Lecture
10	Lab 6	Lecture
11	Lab 7	Lecture
12	Lab 8	Lecture
13	Public Holiday	Quiz 2 (Part 2)

Part A of each lab will have to be completed in class so TAs can help you get started.