EE3801 Data Engineering Principles

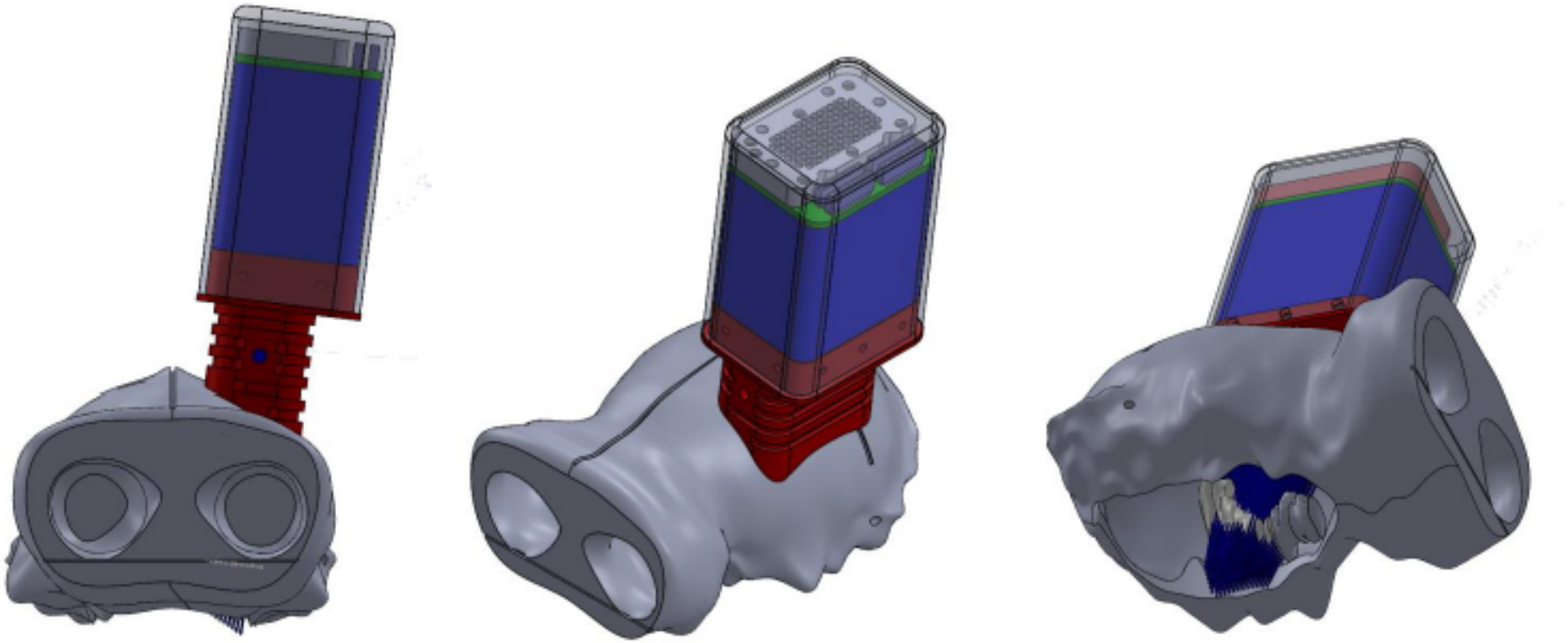
Shih-Cheng YEN

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

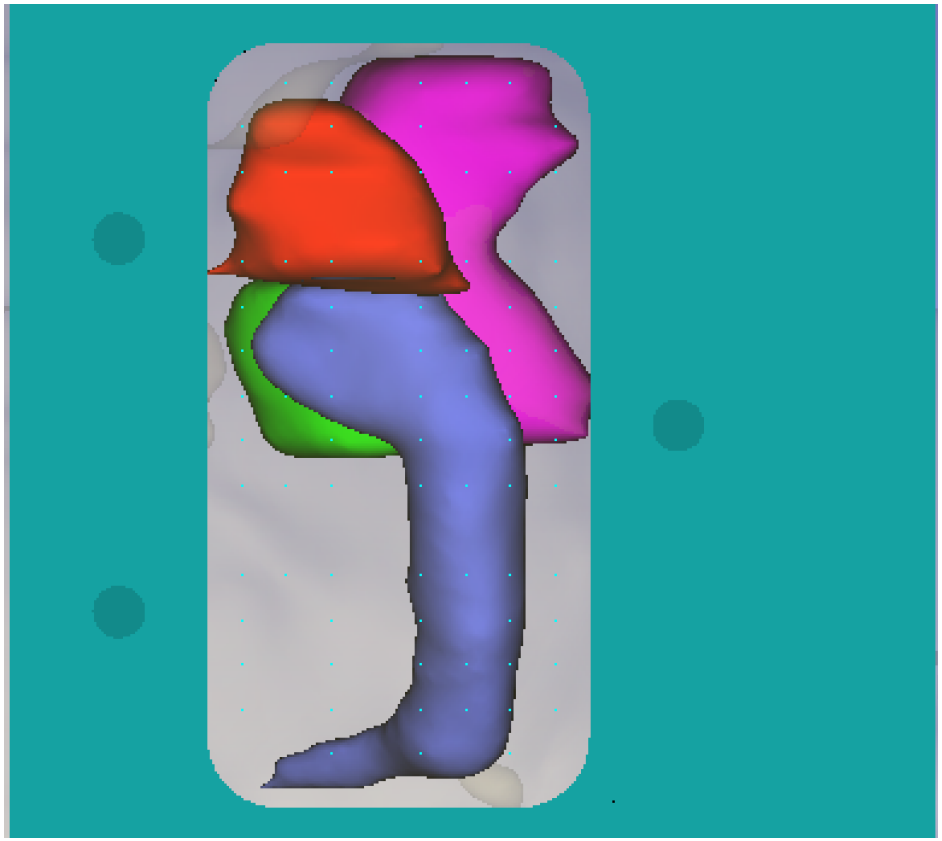
Dataset

Intra-cranial neural recordings with > 100 electrodes



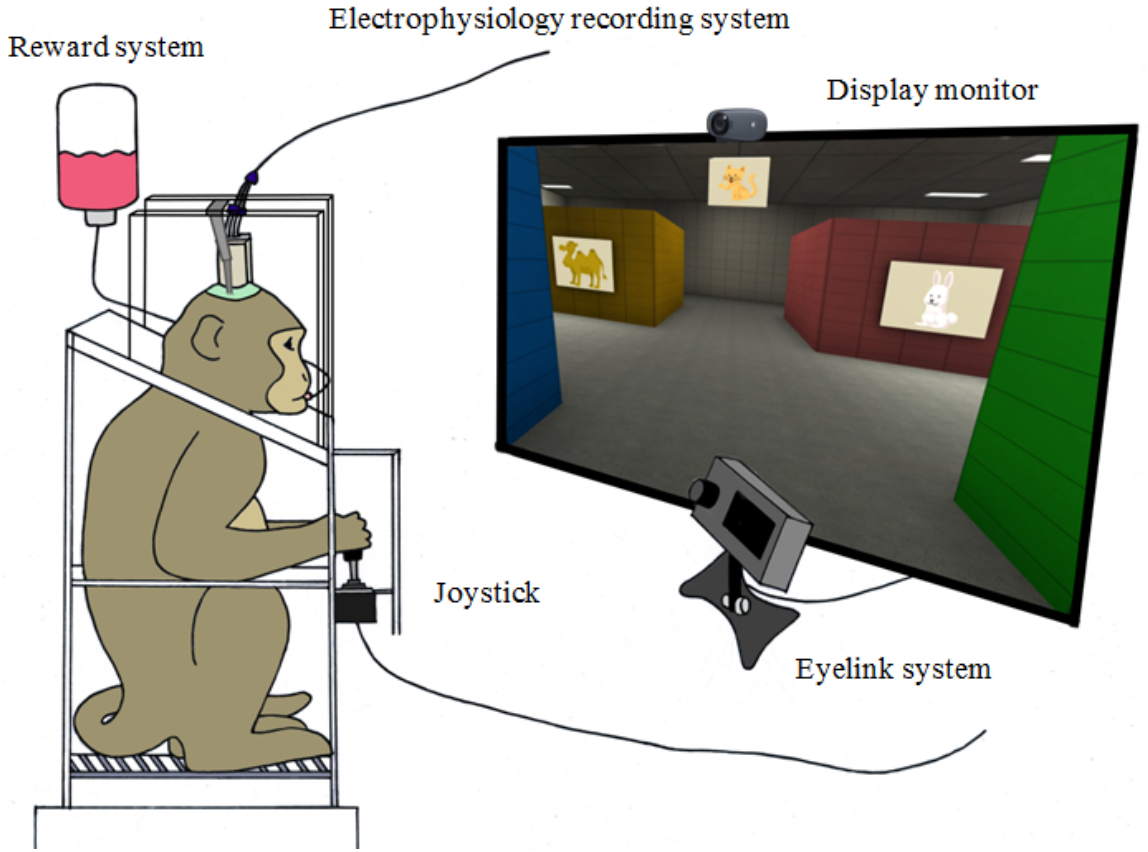
Dataset

Intra-cranial neural recordings with > 100 electrodes



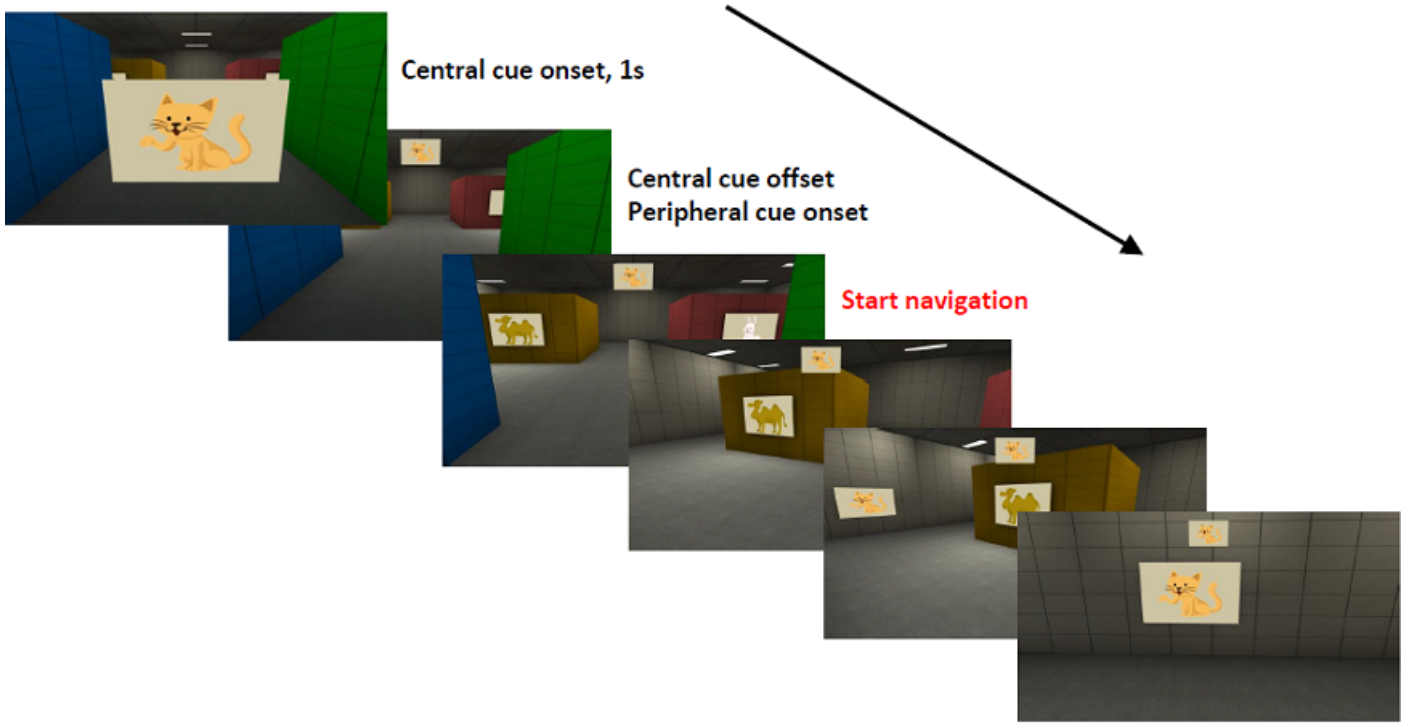
Dataset

Awake behaving animal



Dataset

Experimental task

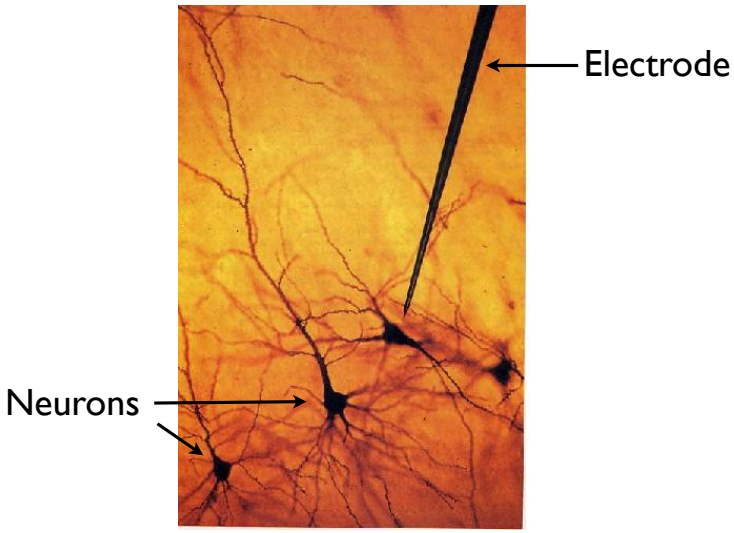


3 possible outcomes:

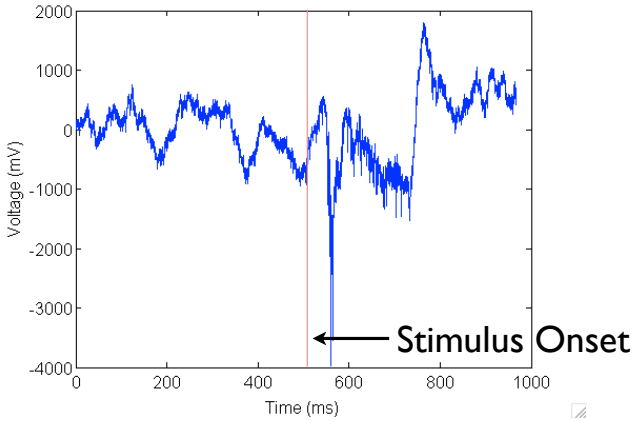
1) Rewarded, cue-target match
2) Error, cue-target mismatch
3) Timeout

Dataset

Neural recordings



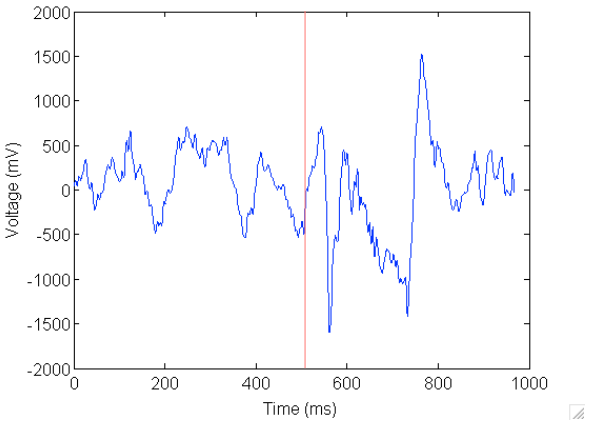
Capable of recording from multiple neurons



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 ≈ 40 GB.

Dataset

Neural recordings

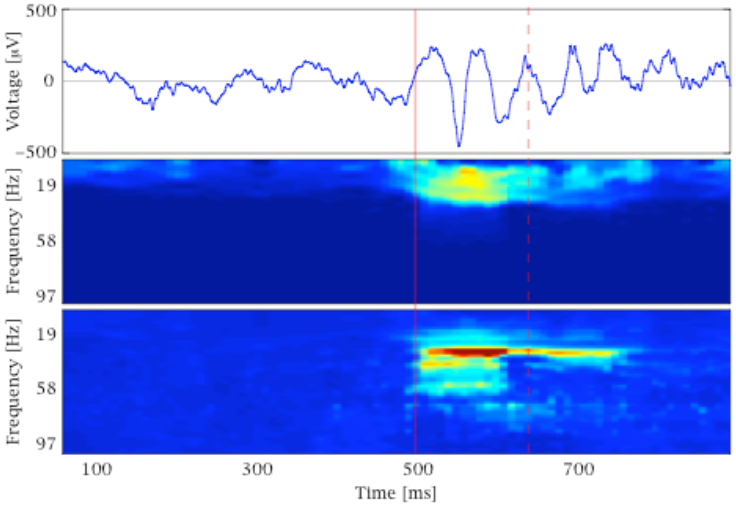


Low-pass filtered at 1 – 150 Hz
Local Field Potential - population synaptic activity

LFP

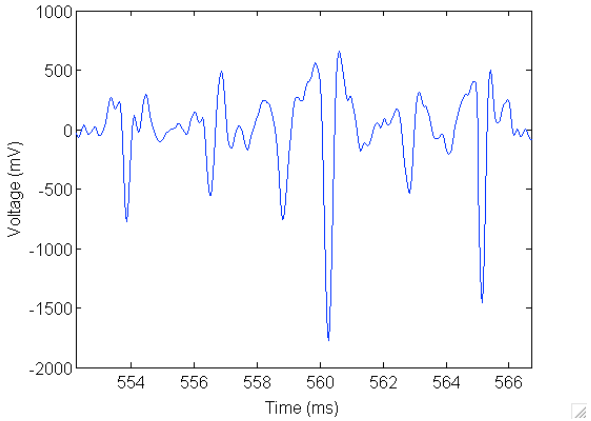
Spectrogram

Normalized Spectrogram

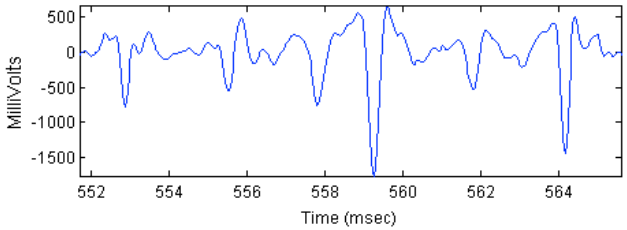


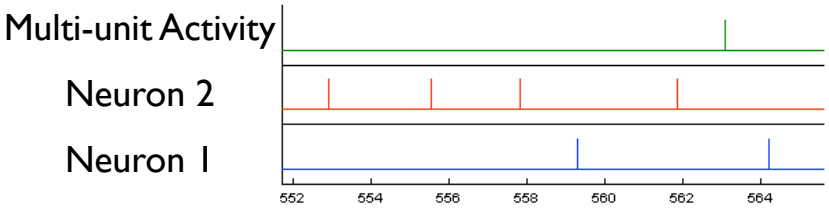
Dataset

Neural recordings



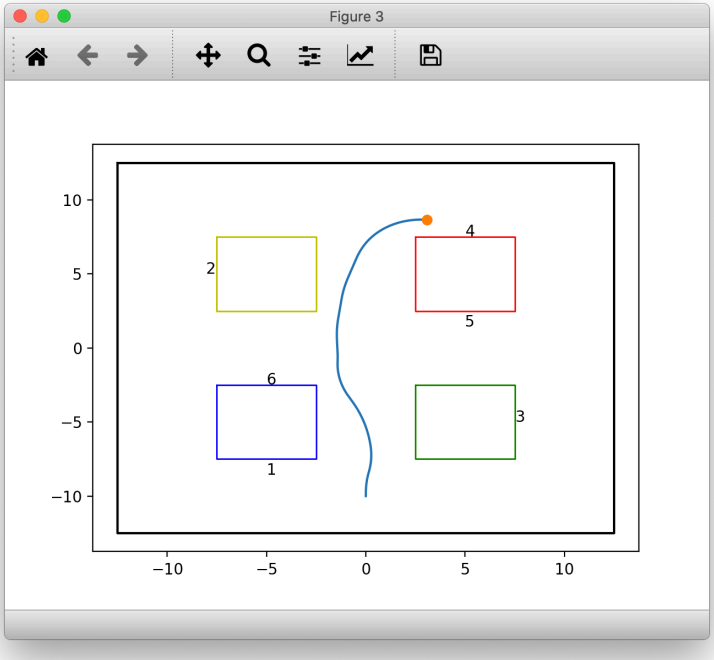
Spikes from multiple neurons present

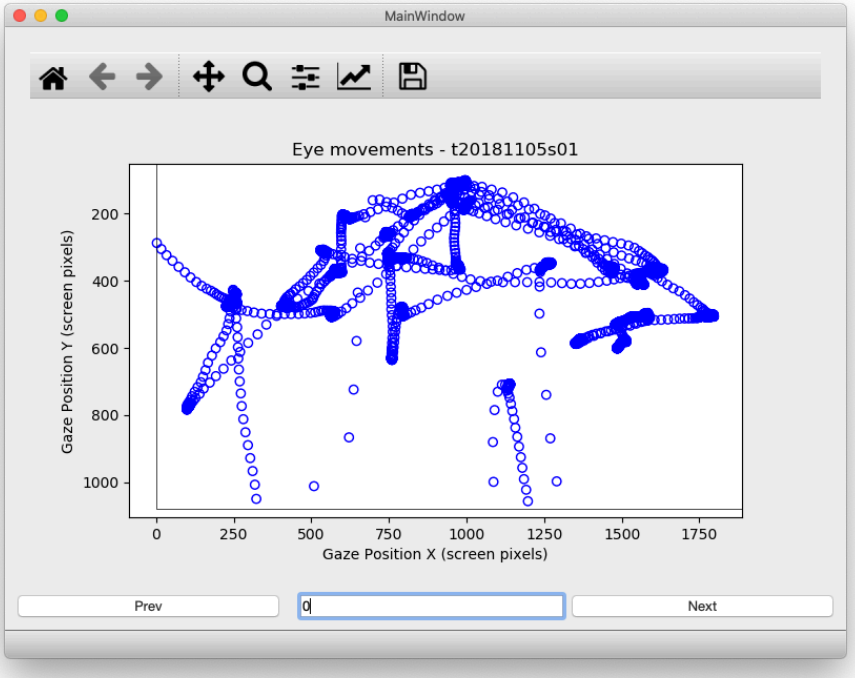




Dataset

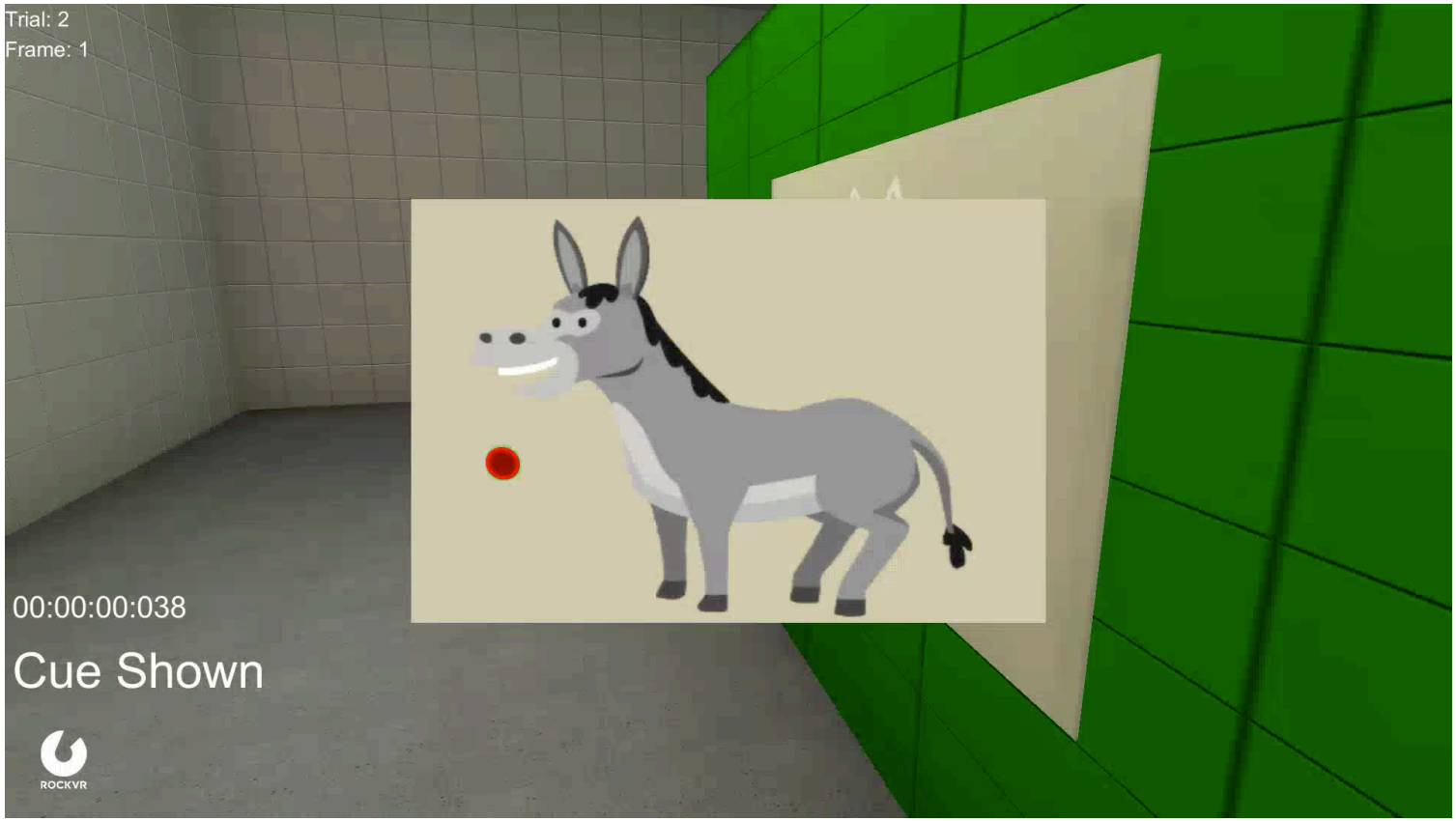
Position in Unity3D maze & eye-position





Dataset

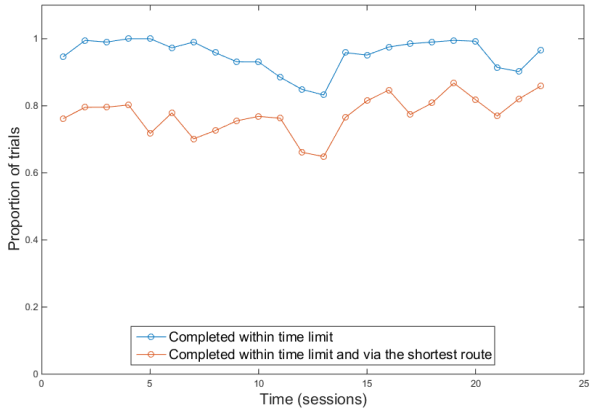
Raycasting



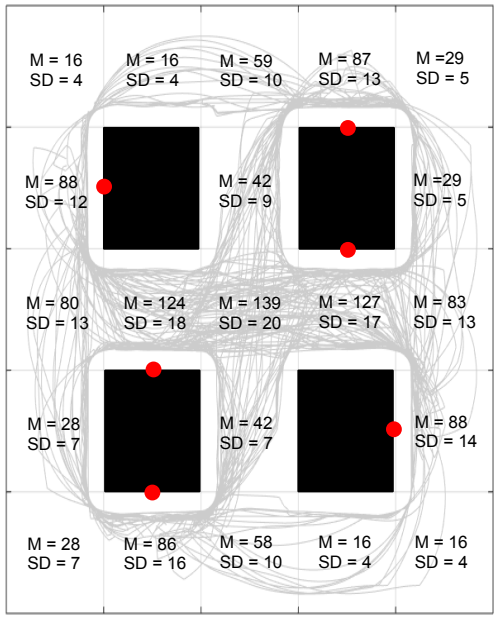
Dataset

Results

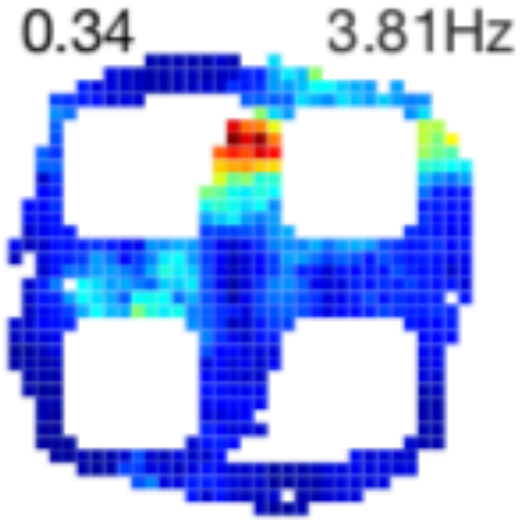
Performance



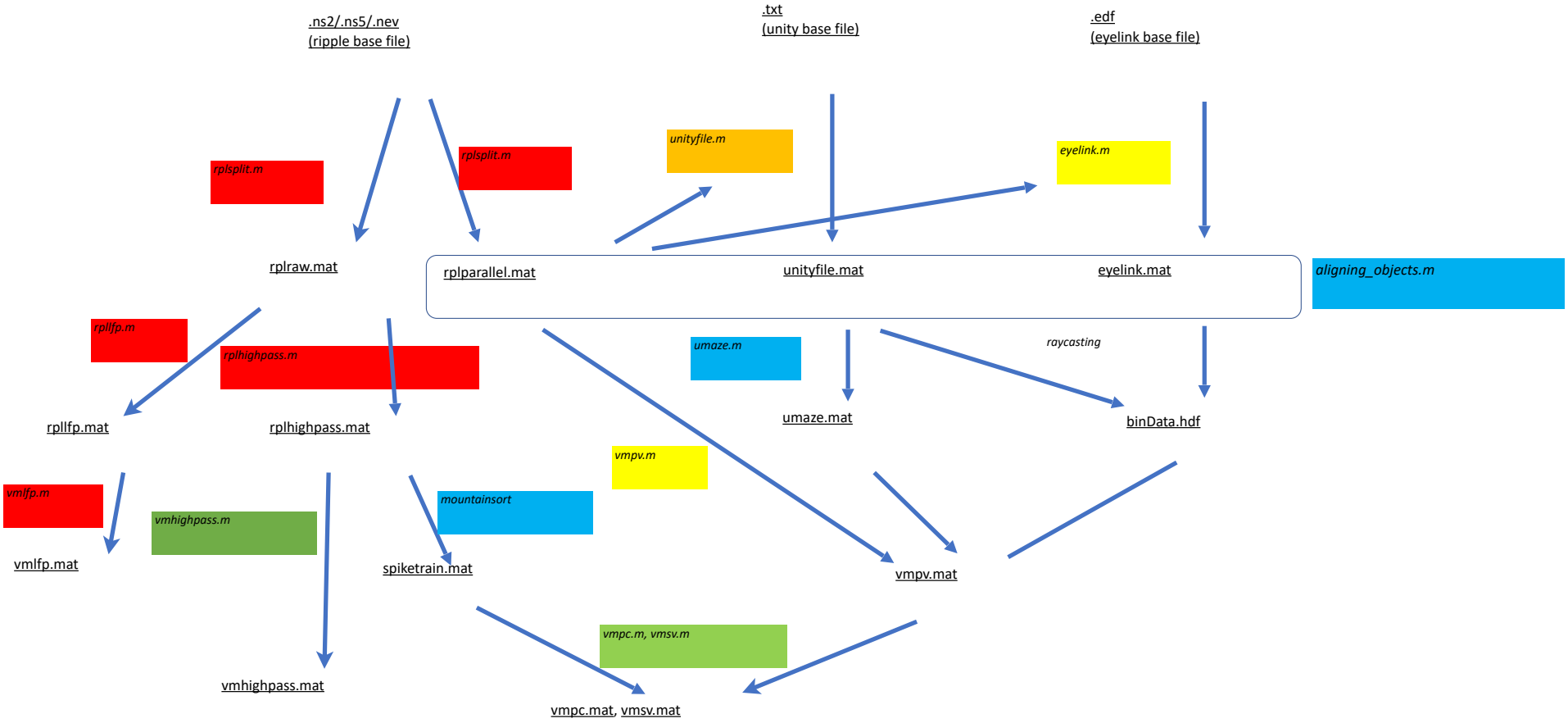
Mean traversals per spatial bin



Place cell



Data pipeline



Parallel Data Processing

Parallel Processing

Serial Pipeline

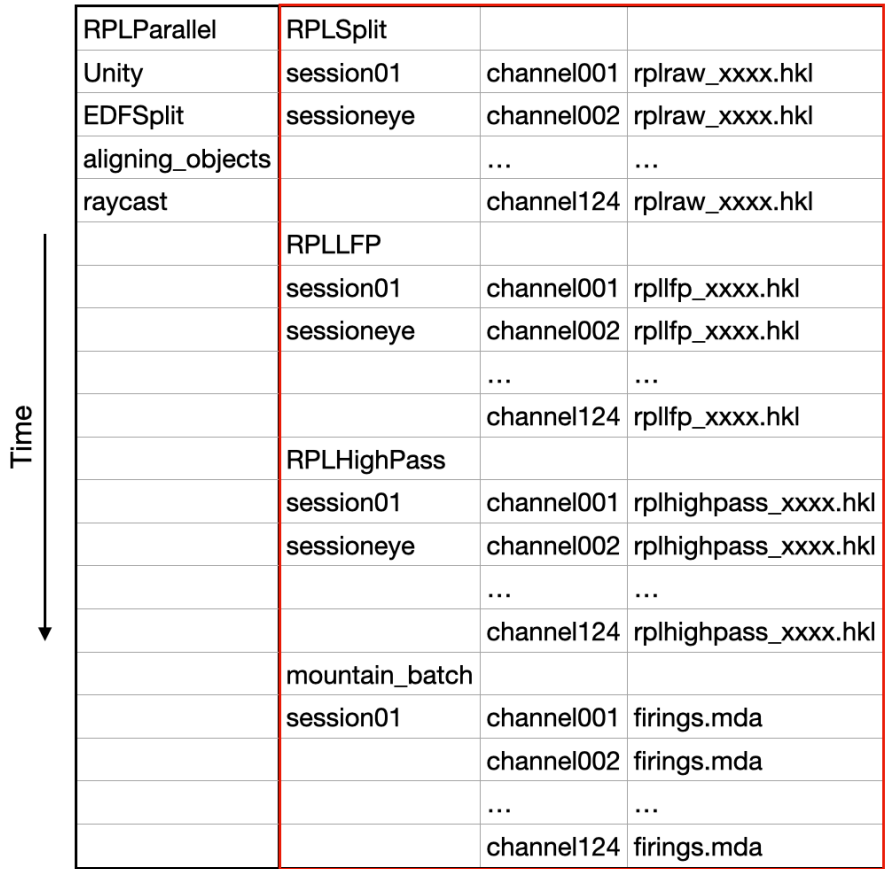
Dependencies	RPLParallel	RPLSplit	RPLLFP	RPLHighPass	Unity	EDFSplit
RPLParallel						
RPLSplit						
RPLLFP		√				
RPLHighPass		√				
Unity	√					
EDFSplit	√					
aligning_objects	√				√	√
raycast					√	√
mountain_batch				√		

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

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
aligning_objects		…		…		…		…
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

Time

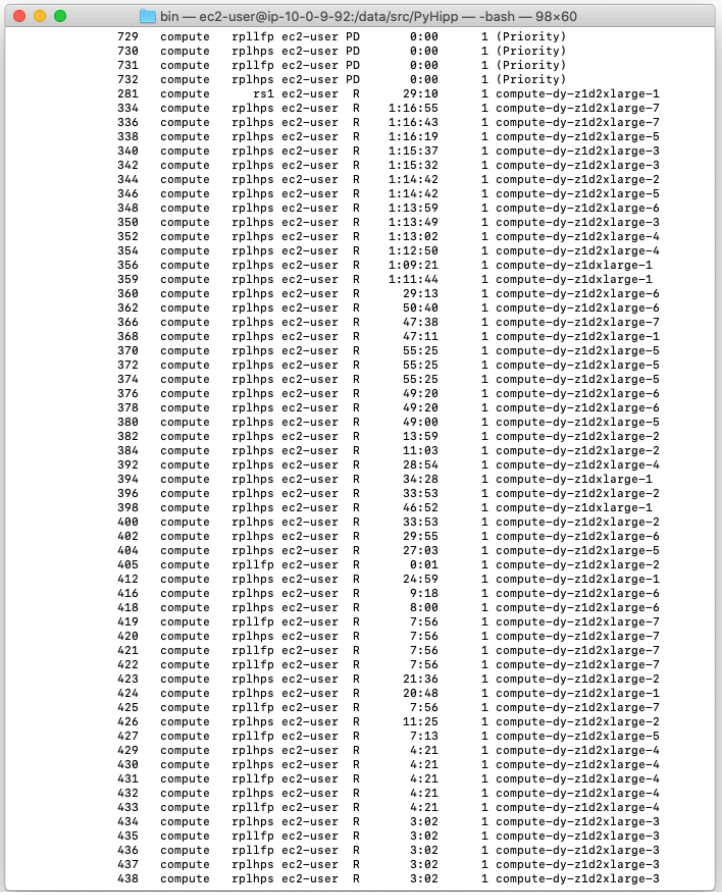
RPLParallel	RPLSplit	
Unity	sessioneye	channel001
EDFSplit		channel002
aligning_objects		…
raycast		channel124
	RPLLFP	
	sessioneye	channel001
		channel002
		…
		channel124
	RPLHighPass	
	sessioneye	channel001
		channel002
		…
		channel124

RPLSplit		RPLLFP	RPLHPS	RPLSplit		RPLLFP	RPLHPS	RPLSplit		RPLLFP	RPLHPS	RPLSplit		RPLLFP	RPLHPS
session01	channel001			session01	channel033			session01	channel065			session01	channel097		
	…				…				…				…		
	channel032				channel064				channel096				channel124		

Jobs: 6 + 110 + 110 = 226

Optimizing Parallel Processing

Fine-Grained Parallel Processing (226 jobs)



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.