Representation learning to improve trace analysis

Quentin Fournier

Dec 6, 2018

Polytechnique Montréal

Laboratoire DORSAL

Contents

Problem

Methods based on trace analysis use representations which lose a lot of information.

Only system call names are considered in most works.

Problem

Complete trace of kernel events

16:04:29.295 148 673	channel0_5	5	syscall_entry_writev	fd=51, vec=140731751376432, vlen=1
16:04:29.295 154 253	channel0_5	5	kmem_cache_alloc_node	call_site=0xffffffff8ac3e93b, ptr=0xffff9c10df8ddd00, bytes_req=256, bytes_alloc=256, gfp_flags=22021312, node=-1
16:04:29.295 155 944	channel0_5	5	kmem_kmalloc_node	call_site=0xffffffff8ac3e967, ptr=0xffff9c0cf3c04800, bytes_req=384, bytes_alloc=512, gfp_flags=22087360, node=-1
16:04:29.295 160 246	channel0_5	5	sched_waking	comm=compiz, tid=2504, prio=20, target_cpu=1
16:04:29.295 163 587	channel0_5	5	sched_migrate_task	comm=compiz, tid=2504, prio=20, orig_cpu=1, dest_cpu=0
16:04:29.295 170 829	channel0_5	5	sched_wakeup	comm=compiz, tid=2504, prio=20, target_cpu=0
16:04:29.295 173 048	channel0_5	5	kmem_kfree	call_site=0xffffffff8a678279, ptr=0x0
16:04:29.295 174 683	channel0_5	5	syscall_exit_writev	ret=32, vec=140731751376432
16:04:29.295 177 782	channel0_5	5	syscall_entry_recvmsg	fd=51, msg=140731751376016, flags=0
16:04:29.295 179 944	channel0_5	5	kmem_kfree	call_site=0xffffffff8ac34d9b, ptr=0x0
16:04:29.295 180 766	channel0_5	5	syscall_exit_recvmsg	ret=-11, msg=140731751376016
16:04:29.295 182 440	channel0_1	1	kmem_kmalloc	call_site=0xffffffffc04d828c, ptr=0xffff9c107d4208c0, bytes_req=48, bytes_alloc=64, gfp_flags=17302048
16:04:29.295 183 461	channel0_5	5	syscall_entry_setitimer	which=0, value=140731751376608
16:04:29.295 184 727	channel0_5	5	timer_hrtimer_cancel	hrtimer=0xffff9c10d58b1088
16:04:29.295 185 644	channel0_5	5	timer_itimer_state	which=0, expires=0, value_sec=0, value_usec=0, interval_sec=0, interval_usec=0
16:04:29.295 186 815	channel0_5	5	syscall_exit_setitimer	ret=0, ovalue=0

Problem

Complete trace of kernel events

16:04:29.295 148 673	channel0_5	5	start writev	fd=51, vec=140731751376432, vlen=1
16:04:29.295 154 253	channel0_5	5	kmem_cache_alloc_node	call_site=0xffffffff8ac3e93b, ptr=0xffff9c10df8ddd00, bytes_req=256, bytes_alloc=256, gfp_flags=22021312, node=-1
16:04:29.295 155 944	channel0_5	5	kmem_kmalloc_node	call_site=0xffffffff8ac3e967, ptr=0xffff9c0cf3c04800, bytes_req=384, bytes_alloc=512, gfp_flags=22087360, node=-1
16:04:29.295 160 246	channel0_5	5	sched_waking	comm=compiz, tid=2504, prio=20, target_cpu=1
16:04:29.295 163 587	channel0_5	5	sched_migrate_task	comm=compiz, tid=2504, prio=20, orig_cpu=1, dest_cpu=0
16:04:29.295 170 829	channel0_5	5	sched_wakeup	comm=compiz, tid=2504, prio=20, target_cpu=0
16:04:29.295 173 048	channel0_5	5	kmem_kfree	call_site=0xffffffff8a678279, ptr=0x0
16:04:29.295 174 683	channel0_5	5	end writev	ret=32, vec=140731751376432
16:04:29.295 177 782	channel0_5	5	syscall_entry_recvmsg	fd=51, msg=140731751376016, flags=0
16:04:29.295 179 944	channel0_5	5	kmem_kfree	call_site=0xffffffff8ac34d9b, ptr=0x0
16:04:29.295 180 766	channel0_5	5	syscall_exit_recvmsg	ret=-11, msg=140731751376016
16:04:29.295 182 440	channel0_1	1	kmem_kmalloc	call_site=0xffffffffc04d828c, ptr=0xffff9c107d4208c0, bytes_req=48, bytes_alloc=64, gfp_flags=17302048
16:04:29.295 183 461	channel0_5	5	syscall_entry_setitimer	which=0, value=140731751376608
16:04:29.295 184 727	channel0_5	5	timer_hrtimer_cancel	hrtimer=0xffff9c10d58b1088
16:04:29.295 185 644	channel0_5	5	timer_itimer_state	which=0, expires=0, value_sec=0, value_usec=0, interval_sec=0, interval_usec=0
16:04:29.295 186 815	channel0_5	5	syscall_exit_setitimer	ret=0, ovalue=0

Problem

Complete trace of kernel events

16:04:29.295 148 673	channel0_5	5	start writev	fd=51, vec=140731751376432, vlen=1
16:04:29.295 154 253	channel0_5	5	kmem_cache_alloc_node	call_site=0xffffffff8ac3e93b, ptr=0xffff9c10df8ddd00, bytes_req=256, bytes_alloc=256, gfp_flags=22021312, node=-1
16:04:29.295 155 944	channel0_5	5	kmem_kmalloc_node	call_site=0xffffffff8ac3e967, ptr=0xffff9c0cf3c04800, bytes_req=384, bytes_alloc=512, gfp_flags=22087360, node=-1
16:04:29.295 160 246	channel0_5	5	sched_waking	comm=compiz, tid=2504, prio=20, target_cpu=1
16:04:29.295 163 587	channel0_5	5	sched_migrate_task	comm=compiz, tid=2504, prio=20, orig_cpu=1, dest_cpu=0
16:04:29.295 170 829	channel0_5	5	sched_wakeup	comm=compiz, tid=2504, prio=20, target_cpu=0
16:04:29.295 173 048	channel0_5	5	kmem_kfree	call_site=0xffffffff8a678279, ptr=0x0
16:04:29.295 174 683	channel0_5	5	end writev	ret=32, vec=140731751376432
16:04:29.295 177 782	channel0_5	5	start recvmsg	fd=51, msg=140731751376016, flags=0
16:04:29.295 179 944	channel0_5	5	kmem_kfree	call_site=0xffffffff8ac34d9b, ptr=0x0
16:04:29.295 180 766	channel0_5	5	end recvmsg	ret=-11, msg=140731751376016
16:04:29.295 182 440	channel0_1	1	kmem_kmalloc	call_site=0xffffffffc04d828c, ptr=0xffff9c107d4208c0, bytes_req=48, bytes_alloc=64, gfp_flags=17302048
16:04:29.295 183 461	channel0_5	5	syscall_entry_setitimer	which=0, value=140731751376608
16:04:29.295 184 727	channel0_5	5	timer_hrtimer_cancel	hrtimer=0xffff9c10d58b1088
16:04:29.295 185 644	channel0_5	5	timer_itimer_state	which=0, expires=0, value_sec=0, value_usec=0, interval_sec=0, interval_usec=0
16:04:29.295 186 815	channel0_5	5	syscall_exit_setitimer	ret=0, ovalue=0

Problem

Complete trace of kernel events

16:04:29.295 148 673	channel0_5	5	start writev	fd=51, vec=140731751376432, vlen=1
16:04:29.295 154 253	channel0_5	5	kmem_cache_alloc_node	call_site=0xffffffff8ac3e93b, ptr=0xffff9c10df8ddd00, bytes_req=256, bytes_alloc=256, gfp_flags=22021312, node=-1
16:04:29.295 155 944	channel0_5	5	kmem_kmalloc_node	call_site=0xffffffff8ac3e967, ptr=0xffff9c0cf3c04800, bytes_req=384, bytes_alloc=512, gfp_flags=22087360, node=-1
16:04:29.295 160 246	channel0_5	5	sched_waking	comm=compiz, tid=2504, prio=20, target_cpu=1
16:04:29.295 163 587	channel0_5	5	sched_migrate_task	comm=compiz, tid=2504, prio=20, orig_cpu=1, dest_cpu=0
16:04:29.295 170 829	channel0_5	5	sched_wakeup	comm=compiz, tid=2504, prio=20, target_cpu=0
16:04:29.295 173 048	channel0_5	5	kmem_kfree	call_site=0xffffffff8a678279, ptr=0x0
16:04:29.295 174 683	channel0_5	5	end writev	ret=32, vec=140731751376432
16:04:29.295 177 782	channel0_5	5	start recvmsg	fd=51, msg=140731751376016, flags=0
16:04:29.295 179 944	channel0_5	5	kmem_kfree	call_site=0xffffffff8ac34d9b, ptr=0x0
16:04:29.295 180 766	channel0_5	5	end recvmsg	ret=-11, msg=140731751376016
16:04:29.295 182 440	channel0_1	1	kmem_kmalloc	call_site=0xffffffffc04d828c, ptr=0xffff9c107d4208c0, bytes_req=48, bytes_alloc=64, gfp_flags=17302048
16:04:29.295 183 461	channel0_5	5	start sititmer	which=0, value=140731751376608
16:04:29.295 184 727	channel0_5	5	timer_hrtimer_cancel	hrtimer=0xffff9c10d58b1088
16:04:29.295 185 644	channel0_5	5	timer_itimer_state	which=0, expires=0, value_sec=0, value_usec=0, interval_sec=0, interval_usec=0
16:04:29.295 186 815	channel0_5	5	end sititmer	ret=0, ovalue=0

Problem

Complete trace of kernel events

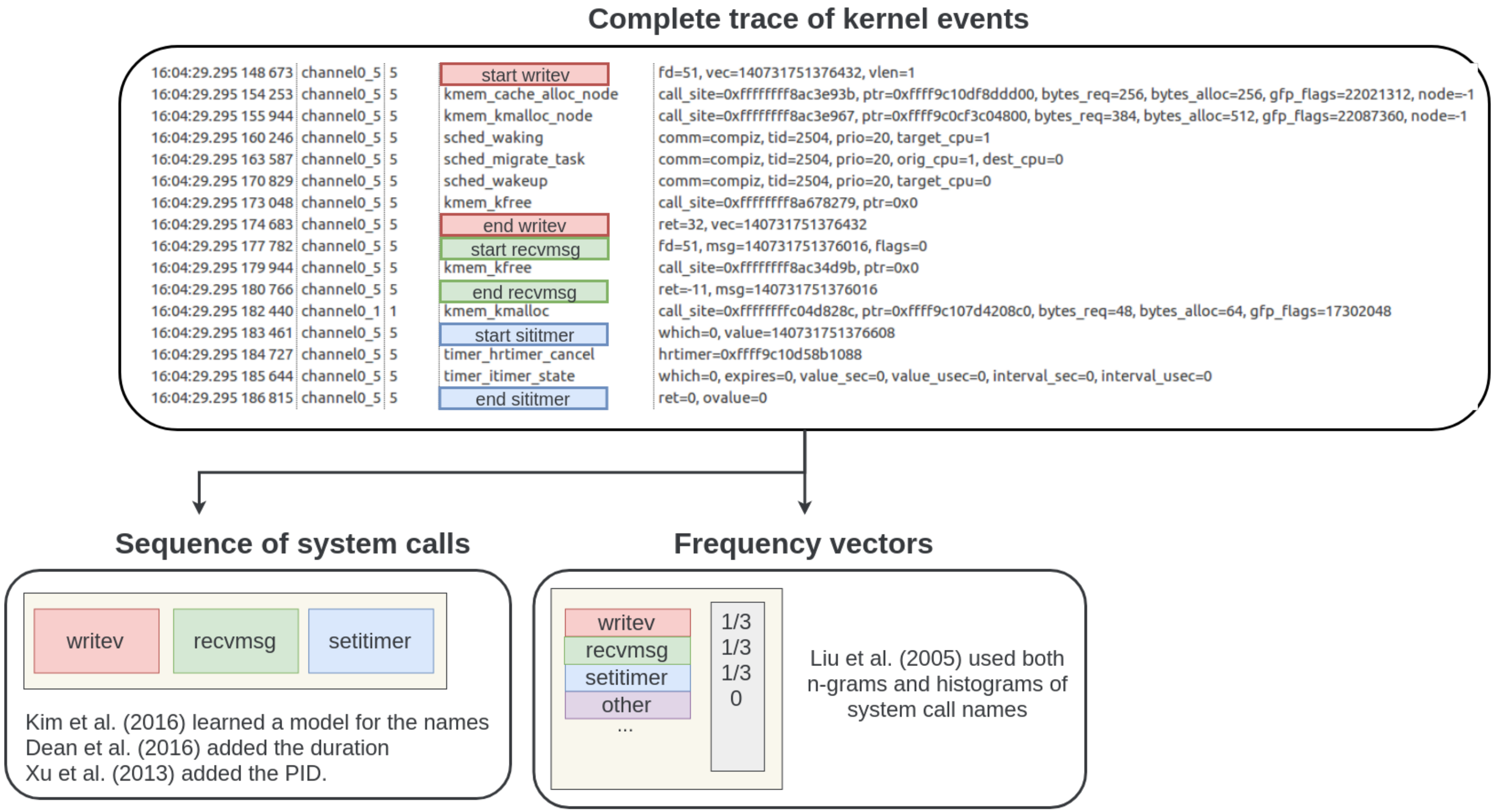
16:04:29.295 148 673	channel0_5	5	start writev	fd=51, vec=140731751376432, vlen=1
16:04:29.295 154 253	channel0_5	5	kmem_cache_alloc_node	call_site=0xffffffff8ac3e93b, ptr=0xffff9c10df8ddd00, bytes_req=256, bytes_alloc=256, gfp_flags=22021312, node=-1
16:04:29.295 155 944	channel0_5	5	kmem_kmalloc_node	call_site=0xffffffff8ac3e967, ptr=0xffff9c0cf3c04800, bytes_req=384, bytes_alloc=512, gfp_flags=22087360, node=-1
16:04:29.295 160 246	channel0_5	5	sched_waking	comm=compiz, tid=2504, prio=20, target_cpu=1
16:04:29.295 163 587	channel0_5	5	sched_migrate_task	comm=compiz, tid=2504, prio=20, orig_cpu=1, dest_cpu=0
16:04:29.295 170 829	channel0_5	5	sched_wakeup	comm=compiz, tid=2504, prio=20, target_cpu=0
16:04:29.295 173 048	channel0_5	5	kmem_kfree	call_site=0xffffffff8a678279, ptr=0x0
16:04:29.295 174 683	channel0_5	5	end writev	ret=32, vec=140731751376432
16:04:29.295 177 782	channel0_5	5	start recvmsg	fd=51, msg=140731751376016, flags=0
16:04:29.295 179 944	channel0_5	5	kmem_kfree	call_site=0xffffffff8ac34d9b, ptr=0x0
16:04:29.295 180 766	channel0_5	5	end recvmsg	ret=-11, msg=140731751376016
16:04:29.295 182 440	channel0_1	1	kmem_kmalloc	call_site=0xffffffffc04d828c, ptr=0xffff9c107d4208c0, bytes_req=48, bytes_alloc=64, gfp_flags=17302048
16:04:29.295 183 461	channel0_5	5	start sititmer	which=0, value=140731751376608
16:04:29.295 184 727	channel0_5	5	timer_hrtimer_cancel	hrtimer=0xffff9c10d58b1088
16:04:29.295 185 644	channel0_5	5	timer_itimer_state	which=0, expires=0, value_sec=0, value_usec=0, interval_sec=0, interval_usec=0
16:04:29.295 186 815	channel0_5	5	end sititmer	ret=0, ovalue=0

Sequence of system calls

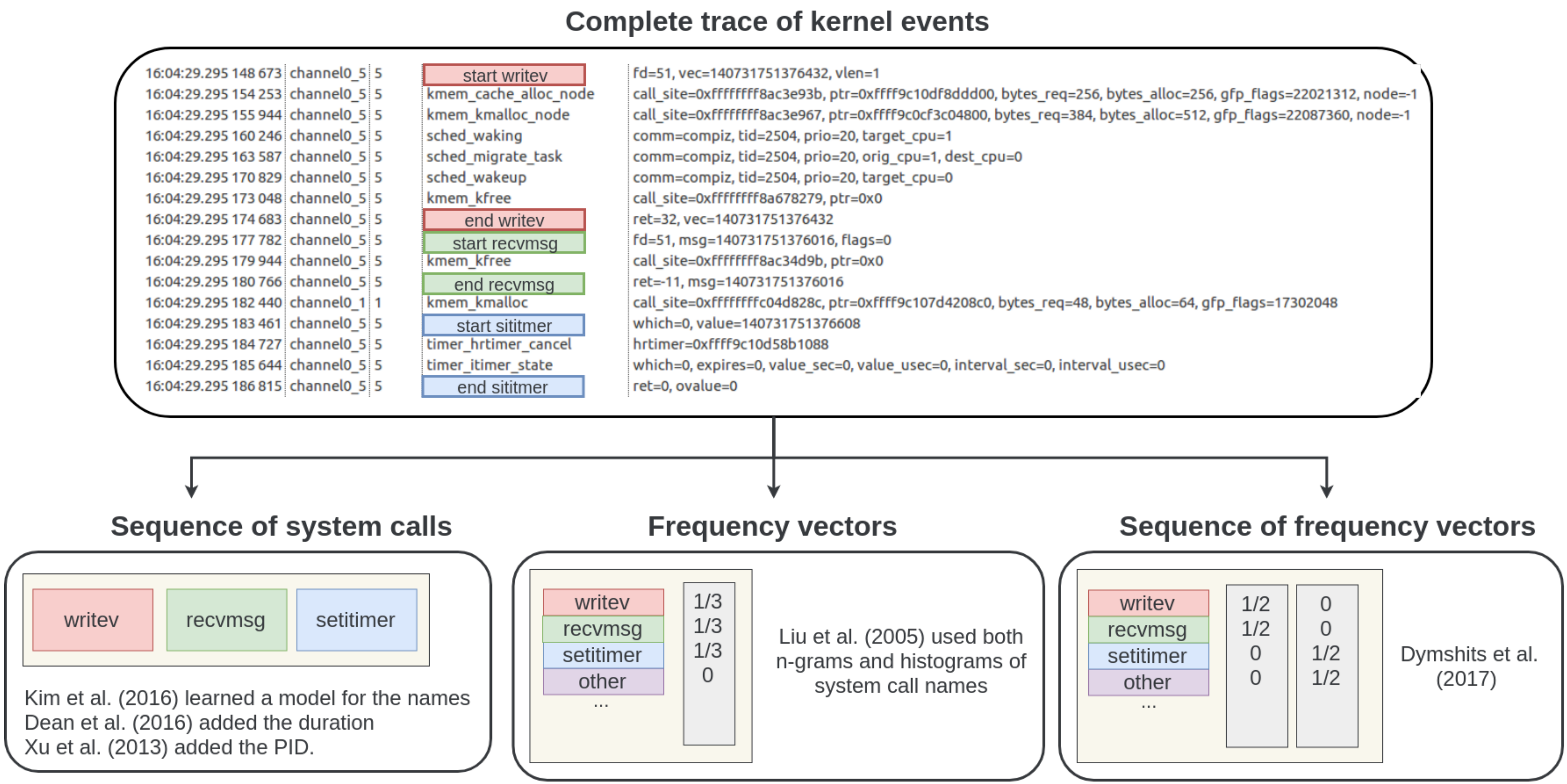
writev | recvmsg | setitimer

Kim et al. (2016) learned a model for the names
Dean et al. (2016) added the duration
Xu et al. (2013) added the PID.

Problem


Complete trace of kernel events
16:04:29.295 148 673 | channel0_5 | 5 | start writev | fd=51, vec=140731751376432, vlen=1
16:04:29.295 154 253 | channel0_5 | 5 | kmem_cache_alloc_node | call_site=0xffffffff8ac3e93b, ptr=0xffff9c10df8ddd00, bytes_req=256, bytes_alloc=256, gfp_flags=22021312, node=-1
16:04:29.295 155 944 | channel0_5 | 5 | kmem_kmalloc_node | call_site=0xffffffff8ac3e967, ptr=0xffff9c0cf3c04800, bytes_req=384, bytes_alloc=512, gfp_flags=22087360, node=-1
16:04:29.295 160 246 | channel0_5 | 5 | sched_waking | comm=compiz, tid=2504, prio=20, target_cpu=1
16:04:29.295 163 587 | channel0_5 | 5 | sched_migrate_task | comm=compiz, tid=2504, prio=20, orig_cpu=1, dest_cpu=0
16:04:29.295 170 829 | channel0_5 | 5 | sched_wakeup | comm=compiz, tid=2504, prio=20, target_cpu=0
16:04:29.295 173 048 | channel0_5 | 5 | kmem_kfree | call_site=0xffffffff8a678279, ptr=0x0
16:04:29.295 174 683 | channel0_5 | 5 | end writev | ret=32, vec=140731751376432
16:04:29.295 177 782 | channel0_5 | 5 | start recvmsg | fd=51, msg=140731751376016, flags=0
16:04:29.295 179 944 | channel0_5 | 5 | kmem_kfree | call_site=0xffffffff8ac34d9b, ptr=0x0
16:04:29.295 180 766 | channel0_5 | 5 | end recvmsg | ret=-11, msg=140731751376016
16:04:29.295 182 440 | channel0_1 | 1 | kmem_kmalloc | call_site=0xffffffffc04d828c, ptr=0xffff9c107d4208c0, bytes_req=48, bytes_alloc=64, gfp_flags=17302048
16:04:29.295 183 461 | channel0_5 | 5 | start sititmer | which=0, value=140731751376608
16:04:29.295 184 727 | channel0_5 | 5 | timer_hrtimer_cancel | hrtimer=0xffff9c10d58b1088
16:04:29.295 185 644 | channel0_5 | 5 | timer_itimer_state | which=0, expires=0, value_sec=0, value_usec=0, interval_sec=0, interval_usec=0
16:04:29.295 186 815 | channel0_5 | 5 | end sititmer | ret=0, ovalue=0
Sequence of system calls
writev
recvmsg
setitimer
Kim et al. (2016) learned a model for the names
Dean et al. (2016) added the duration
Xu et al. (2013) added the PID.
Frequency vectors
writev
recvmsg
setitimer
other
...
1/3
1/3
1/3
0
Liu et al. (2005) used both n-grams and histograms of system call names

Problem


Complete trace of kernel events
16:04:29.295 148 673 | channel0_5 | 5 | start writev | fd=51, vec=140731751376432, vlen=1
16:04:29.295 154 253 | channel0_5 | 5 | kmem_cache_alloc_node | call_site=0xffffffff8ac3e93b, ptr=0xffff9c10df8ddd00, bytes_req=256, bytes_alloc=256, gfp_flags=22021312, node=-1
16:04:29.295 155 944 | channel0_5 | 5 | kmem_kmalloc_node | call_site=0xffffffff8ac3e967, ptr=0xffff9c0cf3c04800, bytes_req=384, bytes_alloc=512, gfp_flags=22087360, node=-1
16:04:29.295 160 246 | channel0_5 | 5 | sched_waking | comm=compiz, tid=2504, prio=20, target_cpu=1
16:04:29.295 163 587 | channel0_5 | 5 | sched_migrate_task | comm=compiz, tid=2504, prio=20, orig_cpu=1, dest_cpu=0
16:04:29.295 170 829 | channel0_5 | 5 | sched_wakeup | comm=compiz, tid=2504, prio=20, target_cpu=0
16:04:29.295 173 048 | channel0_5 | 5 | kmem_kfree | call_site=0xffffffff8a678279, ptr=0x0
16:04:29.295 174 683 | channel0_5 | 5 | end writev | ret=32, vec=140731751376432
16:04:29.295 177 782 | channel0_5 | 5 | start recvmsg | fd=51, msg=140731751376016, flags=0
16:04:29.295 179 944 | channel0_5 | 5 | kmem_kfree | call_site=0xffffffff8ac34d9b, ptr=0x0
16:04:29.295 180 766 | channel0_5 | 5 | end recvmsg | ret=-11, msg=140731751376016
16:04:29.295 182 440 | channel0_1 | 1 | kmem_kmalloc | call_site=0xffffffffc04d828c, ptr=0xffff9c107d4208c0, bytes_req=48, bytes_alloc=64, gfp_flags=17302048
16:04:29.295 183 461 | channel0_5 | 5 | start sititmer | which=0, value=140731751376608
16:04:29.295 184 727 | channel0_5 | 5 | timer_hrtimer_cancel | hrtimer=0xffff9c10d58b1088
16:04:29.295 185 644 | channel0_5 | 5 | timer_itimer_state | which=0, expires=0, value_sec=0, value_usec=0, interval_sec=0, interval_usec=0
16:04:29.295 186 815 | channel0_5 | 5 | end sititmer | ret=0, ovalue=0
Sequence of system calls
writev
recvmsg
setitimer
Kim et al. (2016) learned a model for the names
Dean et al. (2016) added the duration
Xu et al. (2013) added the PID.
Frequency vectors
writev
recvmsg
setitimer
other
...
1/3
1/3
1/3
0
Liu et al. (2005) used both n-grams and histograms of system call names
Sequence of frequency vectors
writev
recvmsg
setitimer
other
...
1/2
1/2
0
0
0
0
1/2
1/2
Dymshits et al. (2017)

Problem

Can we leverage the amount of information available to improve trace analysis ?

Problem

Can we leverage the amount of information available to improve trace analysis ?

Tandon and Chan (2006) and Liu et al. (2005) used with success part of the system call parameters to improve intrusion detection and insider threat detection, respectively.

How to solve it

Complete trace of kernel events

16:04:29.295 148 673	channel0_5	5	kmem_kfree	fd=51, vec=140731751376432, vlen=1
16:04:29.295 154 253	channel0_5	5	block_bio_queue	call_site=0xffffffff8ac3e93b, ptr=0xffff9c10df8ddd00, bytes_req=256, bytes_alloc=256, gfp_flags=22021312, node=-1
16:04:29.295 155 944	channel0_5	5	kmem_kmalloc_node	call_site=0xffffffff8ac3e967, ptr=0xffff9c0cf3c04800, bytes_req=384, bytes_alloc=512, gfp_flags=22087360, node=-1
16:04:29.295 160 246	channel0_5	5	sched_waking	comm=compiz, tid=2504, prio=20, target_cpu=1
16:04:29.295 163 587	channel0_5	5	sched_migrate_task	comm=compiz, tid=2504, prio=20, orig_cpu=1, dest_cpu=0
16:04:29.295 170 829	channel0_5	5	sched_wakeup	comm=compiz, tid=2504, prio=20, target_cpu=0
16:04:29.295 173 048	channel0_5	5	kmem_kfree	call_site=0xffffffff8a678279, ptr=0x0
16:04:29.295 174 683	channel0_5	5	syscall_exit_writev	ret=32, vec=140731751376432
16:04:29.295 177 782	channel0_5	5	syscall_entry_recvmsg	fd=51, msg=140731751376016, flags=0
16:04:29.295 179 944	channel0_5	5	kmem_kfree	call_site=0xffffffff8ac34d9b, ptr=0x0
16:04:29.295 180 766	channel0_5	5	syscall_exit_recvmsg	ret=-11, msg=140731751376016
16:04:29.295 182 440	channel0_1	1	kmem_kmalloc	call_site=0xffffffffc04d828c, ptr=0xffff9c107d4208c0, bytes_req=48, bytes_alloc=64, gfp_flags=17302048
16:04:29.295 183 461	channel0_5	5	syscall_entry_setitimer	which=0, value=140731751376608
16:04:29.295 184 727	channel0_5	5	timer_hrtimer_cancel	hrtimer=0xffff9c10d58b1088
16:04:29.295 185 644	channel0_5	5	timer_itimer_state	which=0, expires=0, value_sec=0, value_usec=0, interval_sec=0, interval_usec=0
16:04:29.295 186 815	channel0_5	5	syscall_exit_setitimer	ret=0, ovalue=0

How to solve it

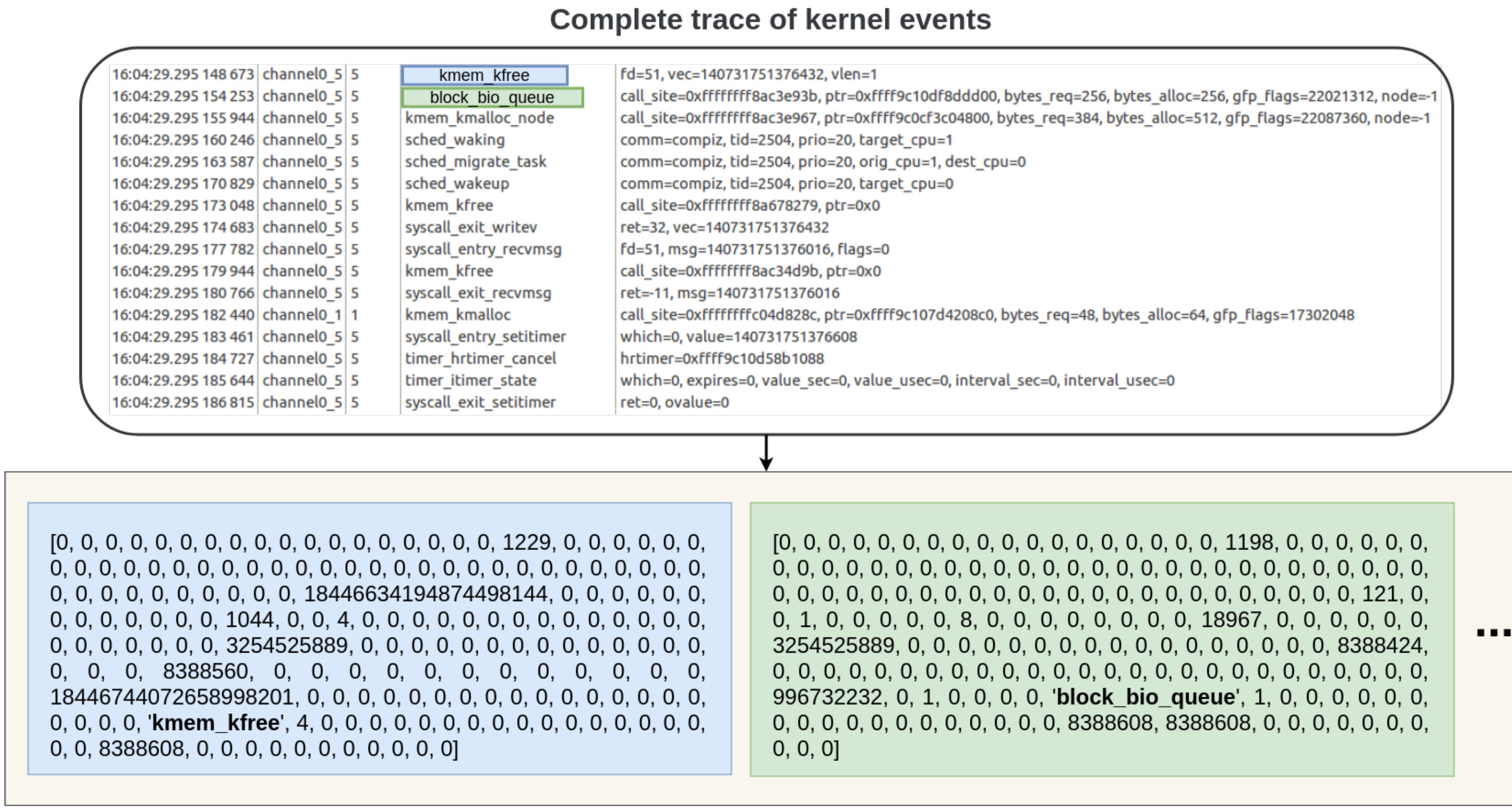
Complete trace of kernel events

16:04:29.295 148 673	channel0_5	5	kmem_kfree	fd=51, vec=140731751376432, vlen=1
16:04:29.295 154 253	channel0_5	5	block_bio_queue	call_site=0xffffffff8ac3e93b, ptr=0xffff9c10df8ddd00, bytes_req=256, bytes_alloc=256, gfp_flags=22021312, node=-1
16:04:29.295 155 944	channel0_5	5	kmem_kmalloc_node	call_site=0xffffffff8ac3e967, ptr=0xffff9c0cf3c04800, bytes_req=384, bytes_alloc=512, gfp_flags=22087360, node=-1
16:04:29.295 160 246	channel0_5	5	sched_waking	comm=compiz, tid=2504, prio=20, target_cpu=1
16:04:29.295 163 587	channel0_5	5	sched_migrate_task	comm=compiz, tid=2504, prio=20, orig_cpu=1, dest_cpu=0
16:04:29.295 170 829	channel0_5	5	sched_wakeup	comm=compiz, tid=2504, prio=20, target_cpu=0
16:04:29.295 173 048	channel0_5	5	kmem_kfree	call_site=0xffffffff8a678279, ptr=0x0
16:04:29.295 174 683	channel0_5	5	syscall_exit_writev	ret=32, vec=140731751376432
16:04:29.295 177 782	channel0_5	5	syscall_entry_recvmsg	fd=51, msg=140731751376016, flags=0
16:04:29.295 179 944	channel0_5	5	kmem_kfree	call_site=0xffffffff8ac34d9b, ptr=0x0
16:04:29.295 180 766	channel0_5	5	syscall_exit_recvmsg	ret=-11, msg=140731751376016
16:04:29.295 182 440	channel0_1	1	kmem_kmalloc	call_site=0xffffffffc04d828c, ptr=0xffff9c107d4208c0, bytes_req=48, bytes_alloc=64, gfp_flags=17302048
16:04:29.295 183 461	channel0_5	5	syscall_entry_setitimer	which=0, value=140731751376608
16:04:29.295 184 727	channel0_5	5	timer_hrtimer_cancel	hrtimer=0xffff9c10d58b1088
16:04:29.295 185 644	channel0_5	5	timer_itimer_state	which=0, expires=0, value_sec=0, value_usec=0, interval_sec=0, interval_usec=0
16:04:29.295 186 815	channel0_5	5	syscall_exit_setitimer	ret=0, ovalue=0

[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1229, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 18446634194874498144, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1044, 0, 0, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3254525889, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 8388560, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 18446744072658998201, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 'kmem_kfree', 4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 8388608, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]

...

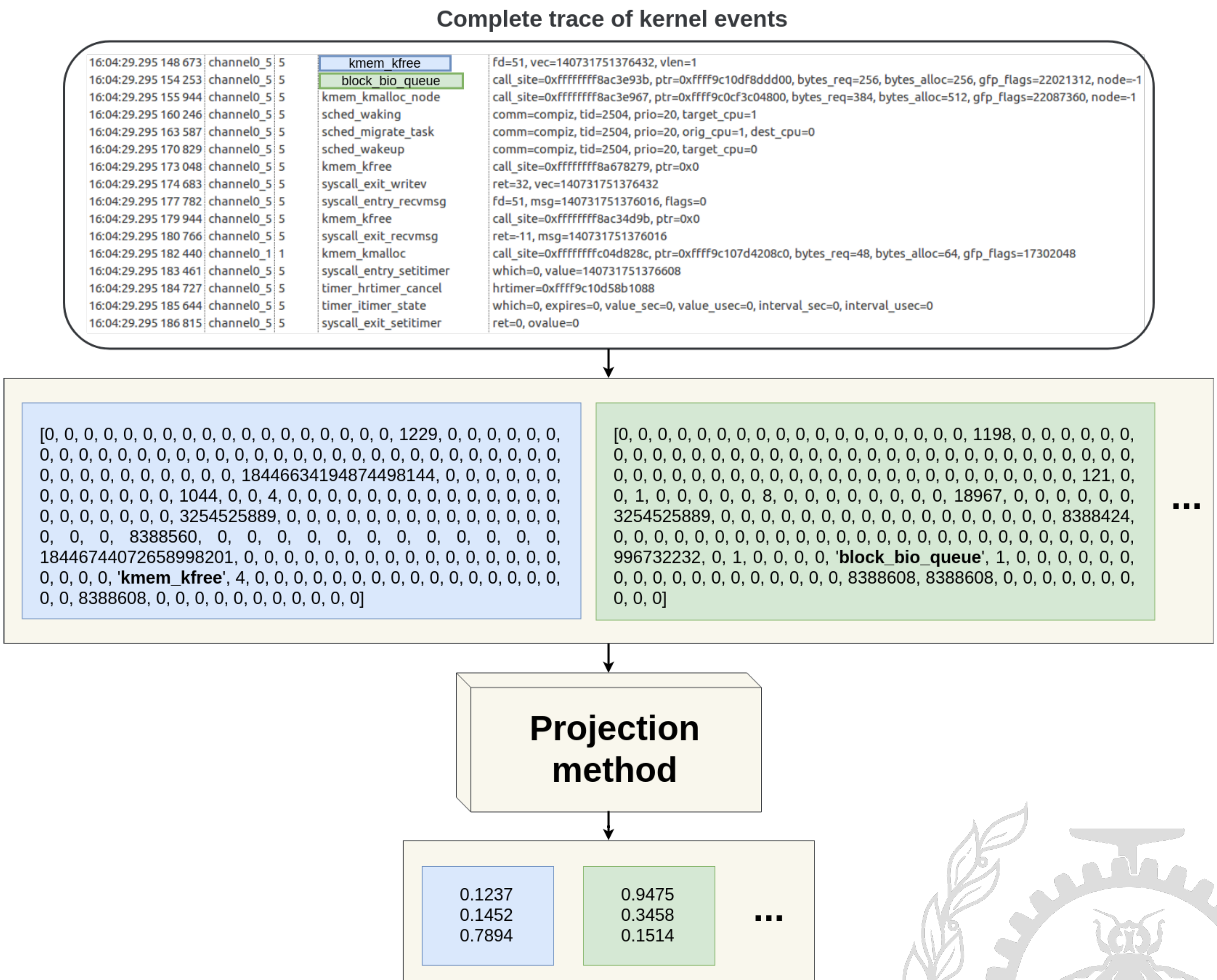
How to solve it



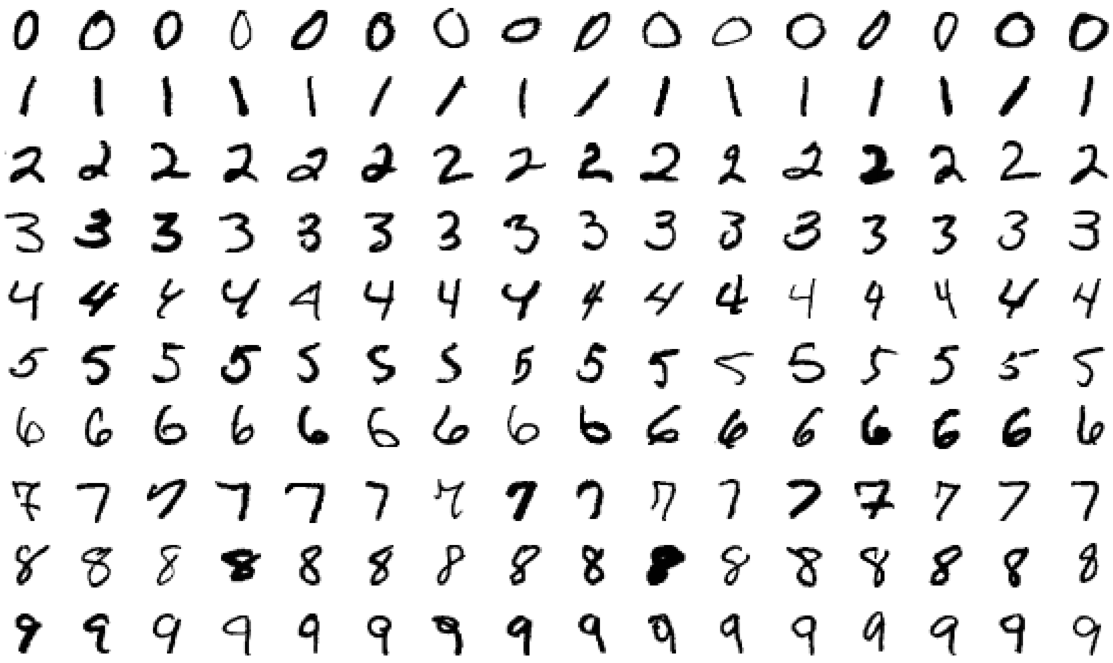
16:04:29.295 148 673	channel0_5	5	kmem_kfree	fd=51, vec=140731751376432, vlen=1
16:04:29.295 154 253	channel0_5	5	block_bio_queue	call_site=0xffffffff8ac3e93b, ptr=0xffff9c10df8ddd00, bytes_req=256, bytes_alloc=256, gfp_flags=22021312, node=-1
16:04:29.295 155 944	channel0_5	5	kmem_kmalloc_node	call_site=0xffffffff8ac3e967, ptr=0xffff9c0cf3c04800, bytes_req=384, bytes_alloc=512, gfp_flags=22087360, node=-1
16:04:29.295 160 246	channel0_5	5	sched_waking	comm=compiz, tid=2504, prio=20, target_cpu=1
16:04:29.295 163 587	channel0_5	5	sched_migrate_task	comm=compiz, tid=2504, prio=20, orig_cpu=1, dest_cpu=0
16:04:29.295 170 829	channel0_5	5	sched_wakeup	comm=compiz, tid=2504, prio=20, target_cpu=0
16:04:29.295 173 048	channel0_5	5	kmem_kfree	call_site=0xffffffff8a678279, ptr=0x0
16:04:29.295 174 683	channel0_5	5	syscall_exit_writev	ret=32, vec=140731751376432
16:04:29.295 177 782	channel0_5	5	syscall_entry_recvmsg	fd=51, msg=140731751376016, flags=0
16:04:29.295 179 944	channel0_5	5	kmem_kfree	call_site=0xffffffff8ac34d9b, ptr=0x0
16:04:29.295 180 766	channel0_5	5	syscall_exit_recvmsg	ret=-11, msg=140731751376016
16:04:29.295 182 440	channel0_1	1	kmem_kmalloc	call_site=0xffffffffc04d828c, ptr=0xffff9c107d4208c0, bytes_req=48, bytes_alloc=64, gfp_flags=17302048
16:04:29.295 183 461	channel0_5	5	syscall_entry_setitimer	which=0, value=140731751376608
16:04:29.295 184 727	channel0_5	5	timer_hrtimer_cancel	hrtimer=0xffff9c10d58b1088
16:04:29.295 185 644	channel0_5	5	timer_itimer_state	which=0, expires=0, value_sec=0, value_usec=0, interval_sec=0, interval_usec=0
16:04:29.295 186 815	channel0_5	5	syscall_exit_setitimer	ret=0, ovalue=0

We need to have a **more compact representation** of events.

How to solve it

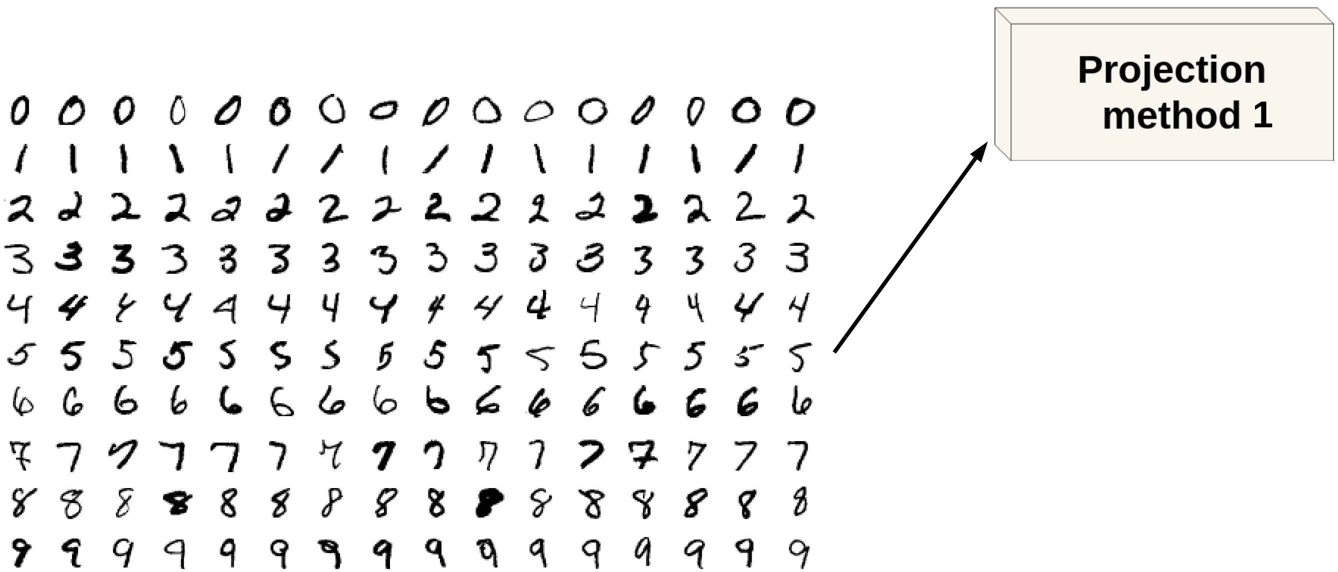


How to solve it



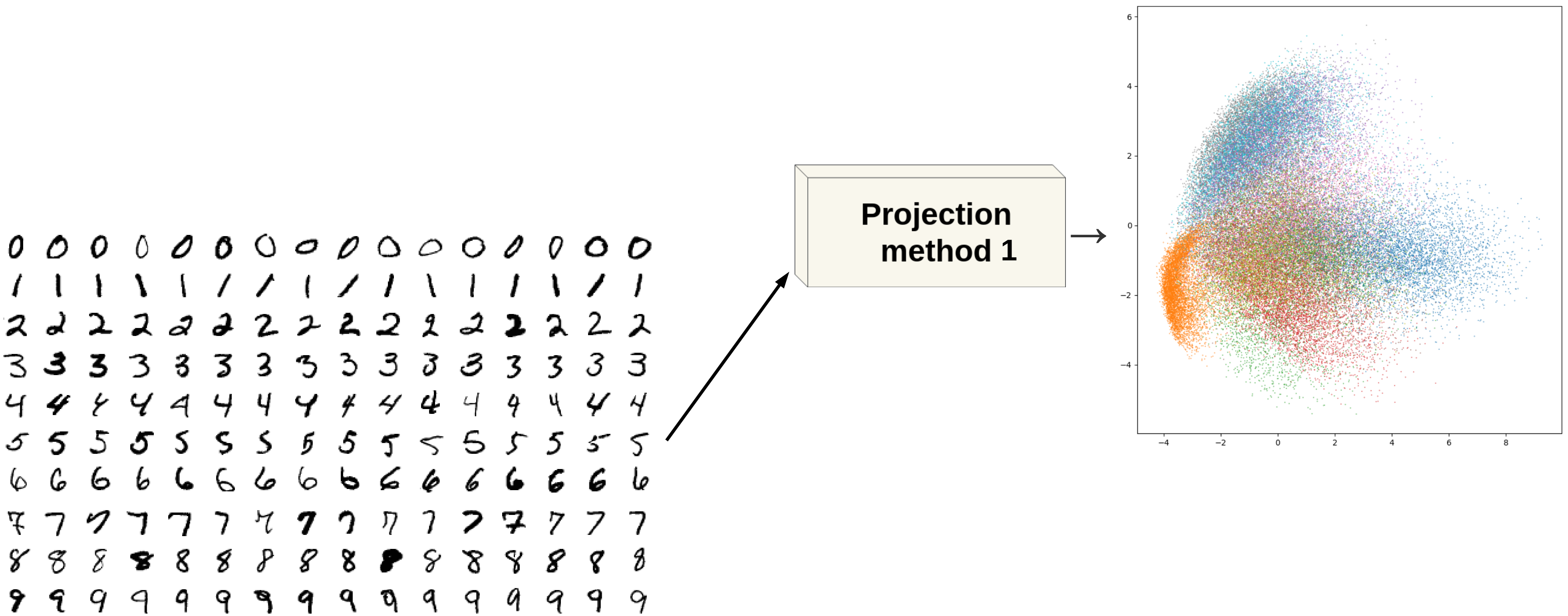
http://yann.lecun.com/exdb/mnist/

How to solve it



http://yann.lecun.com/exdb/mnist/

How to solve it

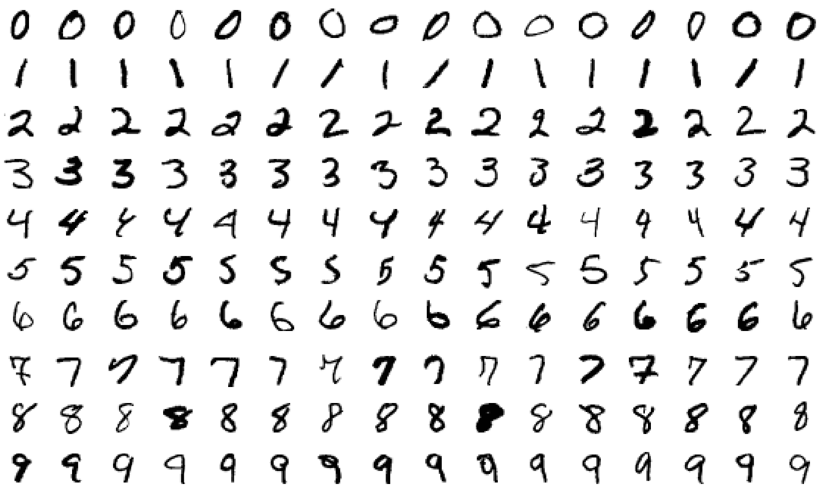


http://yann.lecun.com/exdb/mnist/

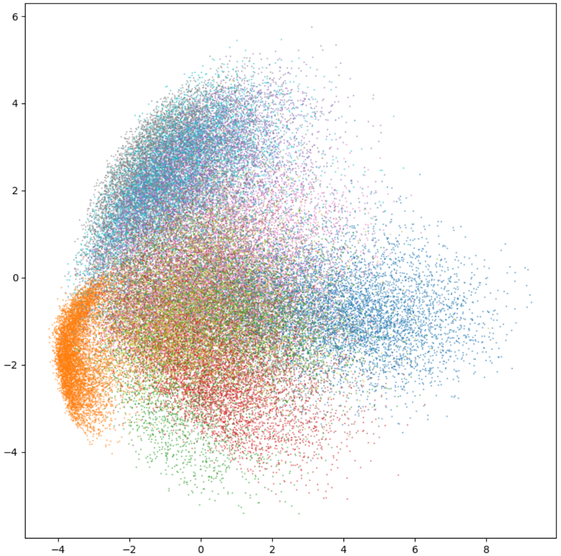
How to solve it

Bad

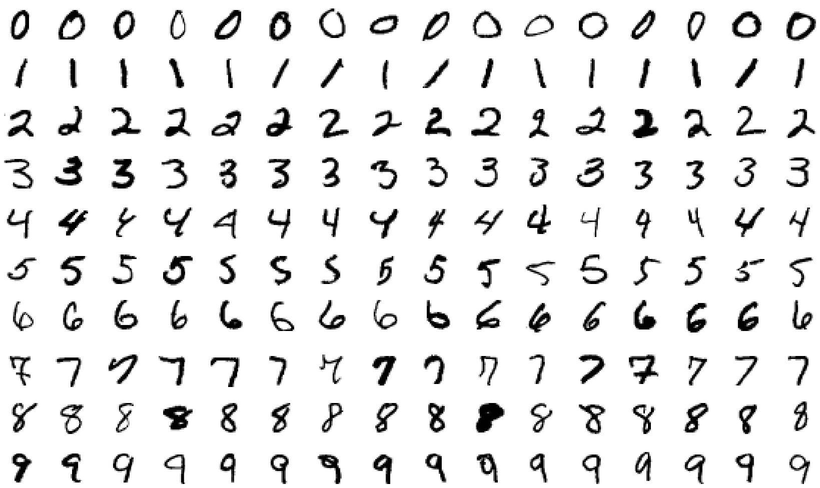
Projection method 1



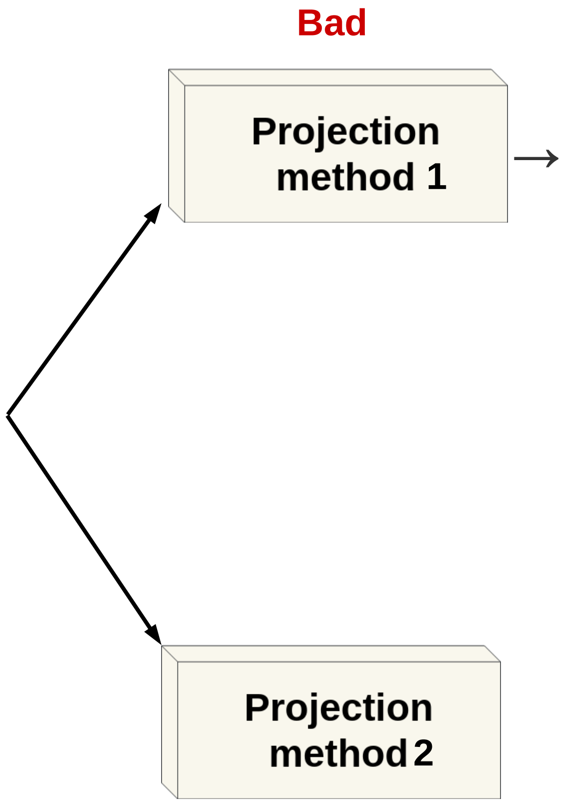
http://yann.lecun.com/exdb/mnist/

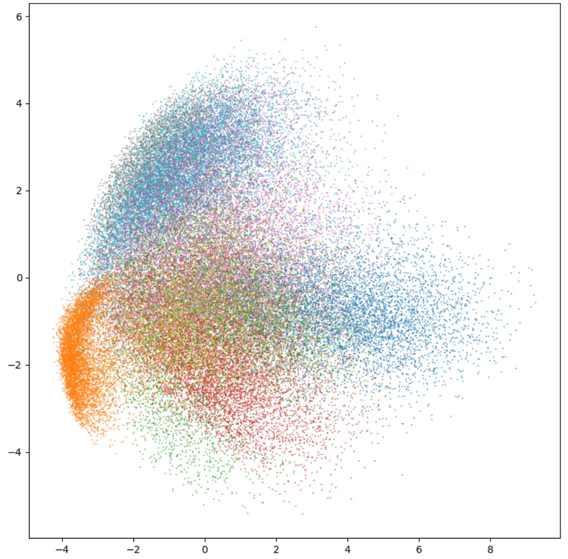


How to solve it



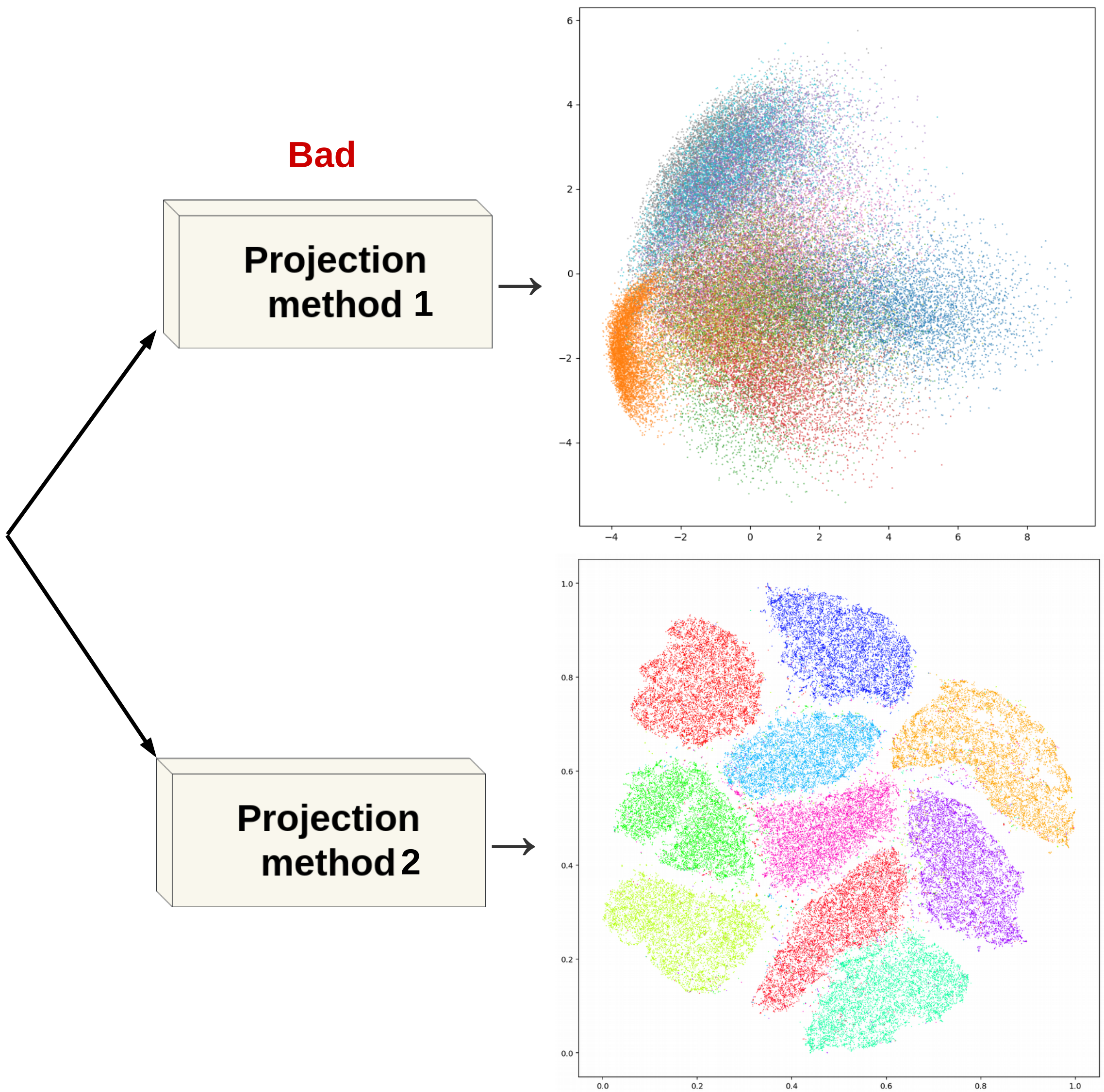
http://yann.lecun.com/exdb/mnist/





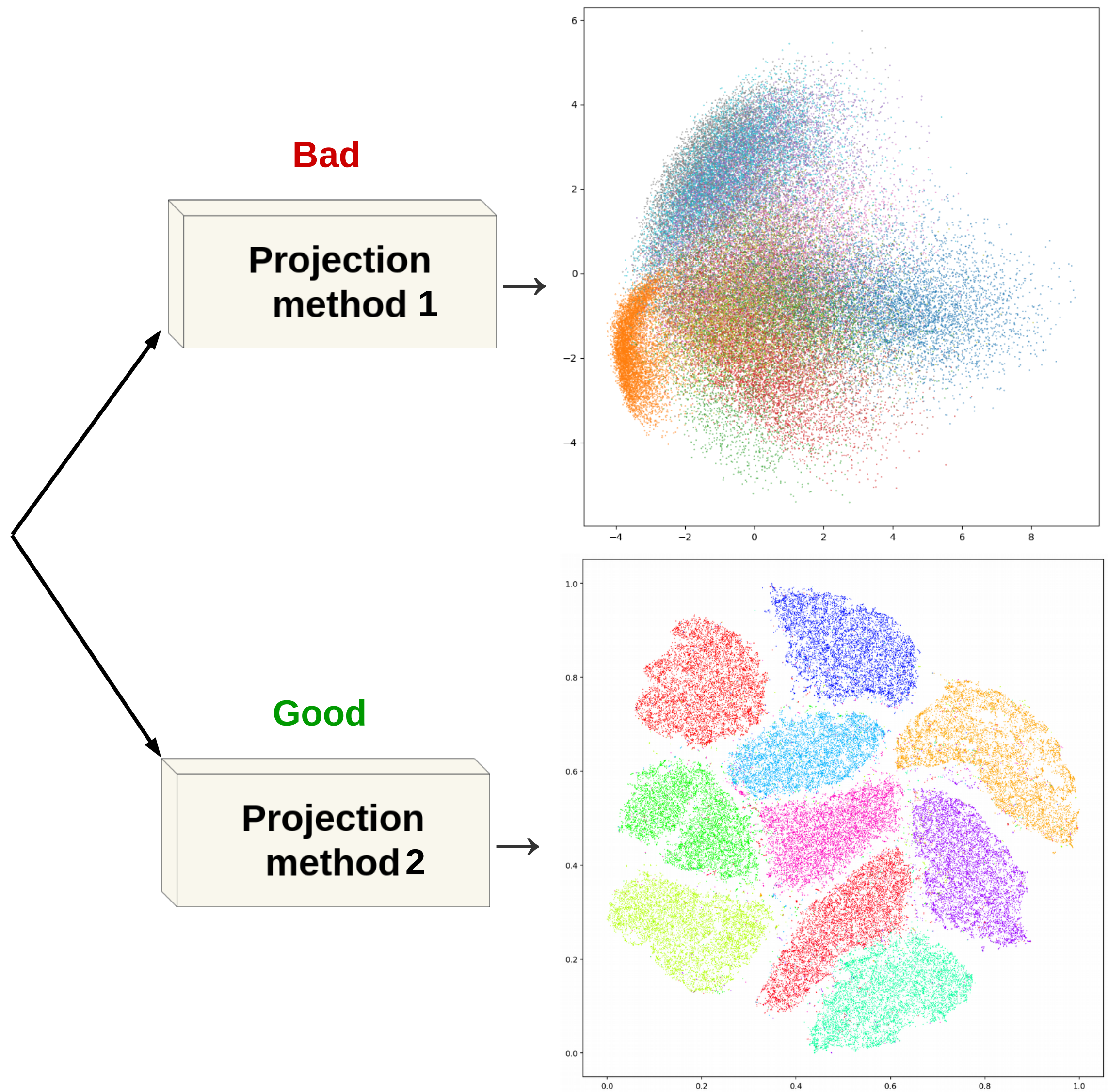
How to solve it

http://yann.lecun.com/exdb/mnist/



How to solve it

http://yann.lecun.com/exdb/mnist/



How to solve it

- No obvious way to represent the data → **Machine Learning**

How to solve it

- No obvious way to represent the data → **Machine Learning**
- Large unlabeled datasets → **Unsupervised Methods**

How to solve it

- No obvious way to represent the data → **Machine Learning**
- Large unlabeled datasets → **Unsupervised Methods**
- Sequence of events → **RNN or Transformer**

How to solve it

Recent success in NLP:

DATASET	TASK	SOTA	OURS
SNLI	Textual Entailment	89.3	**89.9**
MNLI Matched	Textual Entailment	80.6	**82.1**
MNLI Mismatched	Textual Entailment	80.1	**81.4**
SciTail	Textual Entailment	83.3	**88.3**
QNLI	Textual Entailment	82.3	**88.1**
RTE	Textual Entailment	**61.7**	56.0
STS-B	Semantic Similarity	81.0	**82.0**
QQP	Semantic Similarity	66.1	**70.3**
MRPC	Semantic Similarity	**86.0**	82.3
RACE	Reading Comprehension	53.3	**59.0**
ROCStories	Commonsense Reasoning	77.6	**86.5**
COPA	Commonsense Reasoning	71.2	**78.6**
SST-2	Sentiment Analysis	**93.2**	91.3
CoLA	Linguistic Acceptability	35.0	**45.4**
GLUE	Multi Task Benchmark	68.9	**72.8**

https://blog.openai.com/language-unsupervised (Radford et al. (2018)

How to solve it

Recent success in NLP:

DATASET	TASK	SOTA	OURS	
SNLI	Textual Entailment	89.3	**89.9**	
MNLI Matched	Textual Entailment	80.6	**82.1**	
MNLI Mismatched	Textual Entailment	80.1	**81.4**	
SciTail	Textual Entailment	83.3	**88.3**	
QNLI	Textual Entailment	82.3	**88.1**	
RTE	Textual Entailment	**61.7**	56.0	
STS-B	Semantic Similarity	81.0	**82.0**	
QQP	Semantic Similarity	66.1	**70.3**	+6%
MRPC	Semantic Similarity	**86.0**	82.3	
RACE	Reading Comprehension	53.3	**59.0**	+10%
ROCStories	Commonsense Reasoning	77.6	**86.5**	+11%
COPA	Commonsense Reasoning	71.2	**78.6**	+10%
SST-2	Sentiment Analysis	**93.2**	91.3	
CoLA	Linguistic Acceptability	35.0	**45.4**	+30%
GLUE	Multi Task Benchmark	68.9	**72.8**	

https://blog.openai.com/language-unsupervised (Radford et al. (2018)

Representation Learning

Representation learning is the set of methods that learn a projection of the data that facilitate their analysis.

Representation Learning

The **quality** of the representation:

- Improve the **performance**.
- Reduce the computational **time**.

(Bengio et al., 2013).

Methodology

1) **Generate a dataset** ≈ 10^6 - 10^9 kernel events with **all their parameters**.

Traces should be:

a) Heterogeneous.

b) Non-biased.

c) Representative.

Methodology

2) Develop a **representation method** for events or sequence of events.

   a) Transformer (Vaswani et al., 2017).

   b) Multiplicative-LSTM (Kraus et al., 2016).

   c) Autoencoders such as VAE (Kingma et al., 2013).

Methodology

3) Propose a **set of methods to evaluate** the quality of **the representation**:

    a) Intrusion detection on widely use datasets (ADFA-LD, KDD98, UNM-lpr).

    b) Bottlenecking prediction.

    c) Root cause analysis.

4) **Compare** the performance of our method with the **state-of-the-art**.

References

- Dean, D. J., Nguyen, H., Wang, P., Gu, X., Sailer, A., & Kochut, A. (2016). PerfCompass: Online Performance Anomaly Fault Localization and Inference in Infrastructure-as-a-Service Clouds. IEEE Transactions on Parallel and Distributed Systems, 27(6), 1742–1755.
- Kim, G., Yi, H., Lee, J., Paek, Y., & Yoon, S. (2016). LSTM-Based System-Call Language Modeling and Robust Ensemble Method for Designing Host-Based Intrusion Detection Systems. CoRR, abs/1611.0.
- Xu, Z., Yu, X., Feng, Y., Hu, J., Tari, Z., & Han, F. (2013). A multi-module anomaly detection scheme based on system call prediction. In 2013 IEEE 8th Conference on Industrial Electronics and Applications (ICIEA) (pp. 1376–1381).
- Liu, A., Martin, C., Hetherington, T., & Matzner, S. (2005). A comparison of system call feature representations for insider threat detection. In Proceedings from the Sixth Annual IEEE SMC Information Assurance Workshop (pp. 340–347).
- Dymshits, M., Myara, B., & Tolpin, D. (2017). Process Monitoring on Sequences of System Call Count Vectors. CoRR, abs/1707.0.
- Tandon, G., & Chan, P. K. (2006). On The Learning Of System Call Attributes For Host-based Anomaly Detection. International Journal on Artificial Intelligence Tools, 15(06), 875–892.
- Bengio, Y., Courville, A., & Vincent, P. (2013). Representation Learning: A Review and New Perspectives. IEEE Trans. Pattern Anal. Mach. Intell., 35(8), 1798–1828.
- Radford, A., Józefowicz, R., & Sutskever, I. (2017). Learning to Generate Reviews and Discovering Sentiment. CoRR, abs/1704.0.
- Radford, A. (2018). Improving Language Understanding by Generative Pre-Training.
- Krause, B., Lu, L., Murray, I., & Renals, S. (2016). Multiplicative LSTM for sequence modelling. CoRR, abs/1609.0.
- Kingma, D. P., & Welling, M. (2013). Auto-Encoding Variational Bayes. CoRR, abs/1312.6.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is All you Need. In I. Guyon, U. V Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), Advances in Neural Information Processing Systems 30 (pp. 5998–6008). Curran Associates, Inc.

Questions?