Data Science as a Means to Expedite Software Behavior Analysis

Presented by Joely Nelson

Background

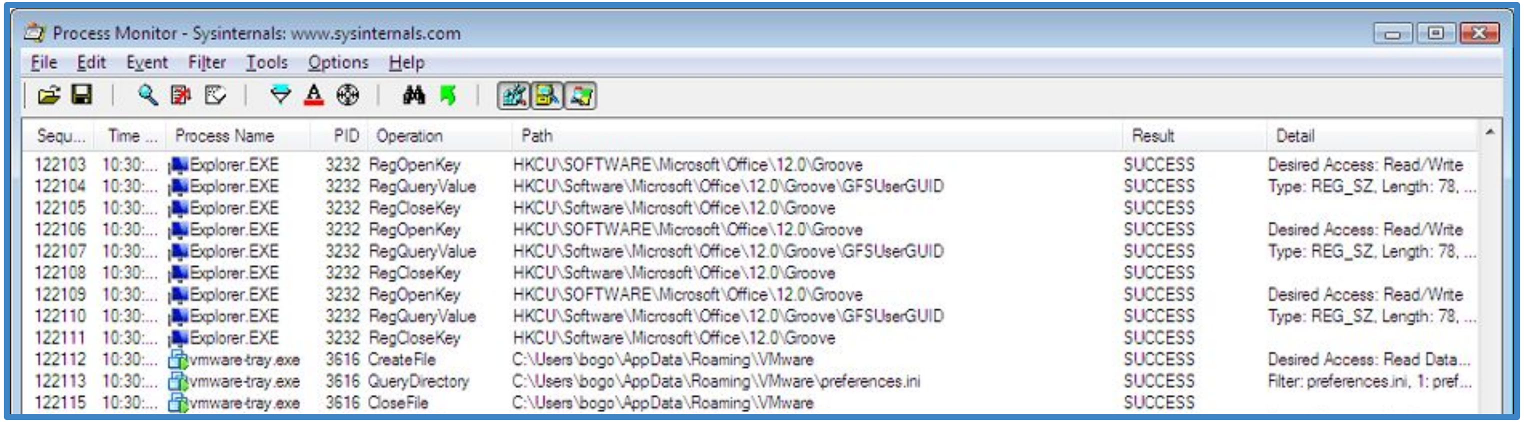
- Joely Nelson (she/her)
- Interests
  - Data science and machine learning applied to domains that have a positive social impact
- Education
  - Received BS of Computer Science with a minor in mathematics from the University of Washington in 2020
  - Will complete MS of Computer Science & Engineering at the University of Washington in March 2022
- Work + Research
  - Computational synthetic biology research
  - R&D Data Science Intern for the Center of Cyber Defenders at Sandia National Laboratories

Research Question: Can we determine software behavior strictly from log analysis?

Motivation

- Programs generate event logs
- These logs can be analyzed manually to determine the behavior of the program

Note: All data in this presentation is synthetic, but is representative of the real data and results.


Process Monitor - Sysinternals: www.sysinternals.com
File Edit Event Filter Tools Options Help

Sequ...	Time ...	Process Name	PID	Operation	Path	Result	Detail
122103	10:30:...	Explorer.EXE	3232	RegOpenKey	HKCU\SOFTWARE\Microsoft\Office\12.0\Groove	SUCCESS	Desired Access: Read/Write
122104	10:30:...	Explorer.EXE	3232	RegQueryValue	HKCU\Software\Microsoft\Office\12.0\Groove\GFSUserGUID	SUCCESS	Type: REG_SZ, Length: 78, ...
122105	10:30:...	Explorer.EXE	3232	RegCloseKey	HKCU\Software\Microsoft\Office\12.0\Groove	SUCCESS	
122106	10:30:...	Explorer.EXE	3232	RegOpenKey	HKCU\SOFTWARE\Microsoft\Office\12.0\Groove	SUCCESS	Desired Access: Read/Write
122107	10:30:...	Explorer.EXE	3232	RegQueryValue	HKCU\Software\Microsoft\Office\12.0\Groove\GFSUserGUID	SUCCESS	Type: REG_SZ, Length: 78, ...
122108	10:30:...	Explorer.EXE	3232	RegCloseKey	HKCU\Software\Microsoft\Office\12.0\Groove	SUCCESS	
122109	10:30:...	Explorer.EXE	3232	RegOpenKey	HKCU\SOFTWARE\Microsoft\Office\12.0\Groove	SUCCESS	Desired Access: Read/Write
122110	10:30:...	Explorer.EXE	3232	RegQueryValue	HKCU\Software\Microsoft\Office\12.0\Groove\GFSUserGUID	SUCCESS	Type: REG_SZ, Length: 78, ...
122111	10:30:...	Explorer.EXE	3232	RegCloseKey	HKCU\Software\Microsoft\Office\12.0\Groove	SUCCESS	
122112	10:30:...	vmware-tray.exe	3616	CreateFile	C:\Users\bogo\AppData\Roaming\VMware	SUCCESS	Desired Access: Read Data...
122113	10:30:...	vmware-tray.exe	3616	QueryDirectory	C:\Users\bogo\AppData\Roaming\VMware\preferences.ini	SUCCESS	Filter: preferences.ini, 1: pref...
122115	10:30:...	vmware-tray.exe	3616	CloseFile	C:\Users\bogo\AppData\Roaming\VMware	SUCCESS	

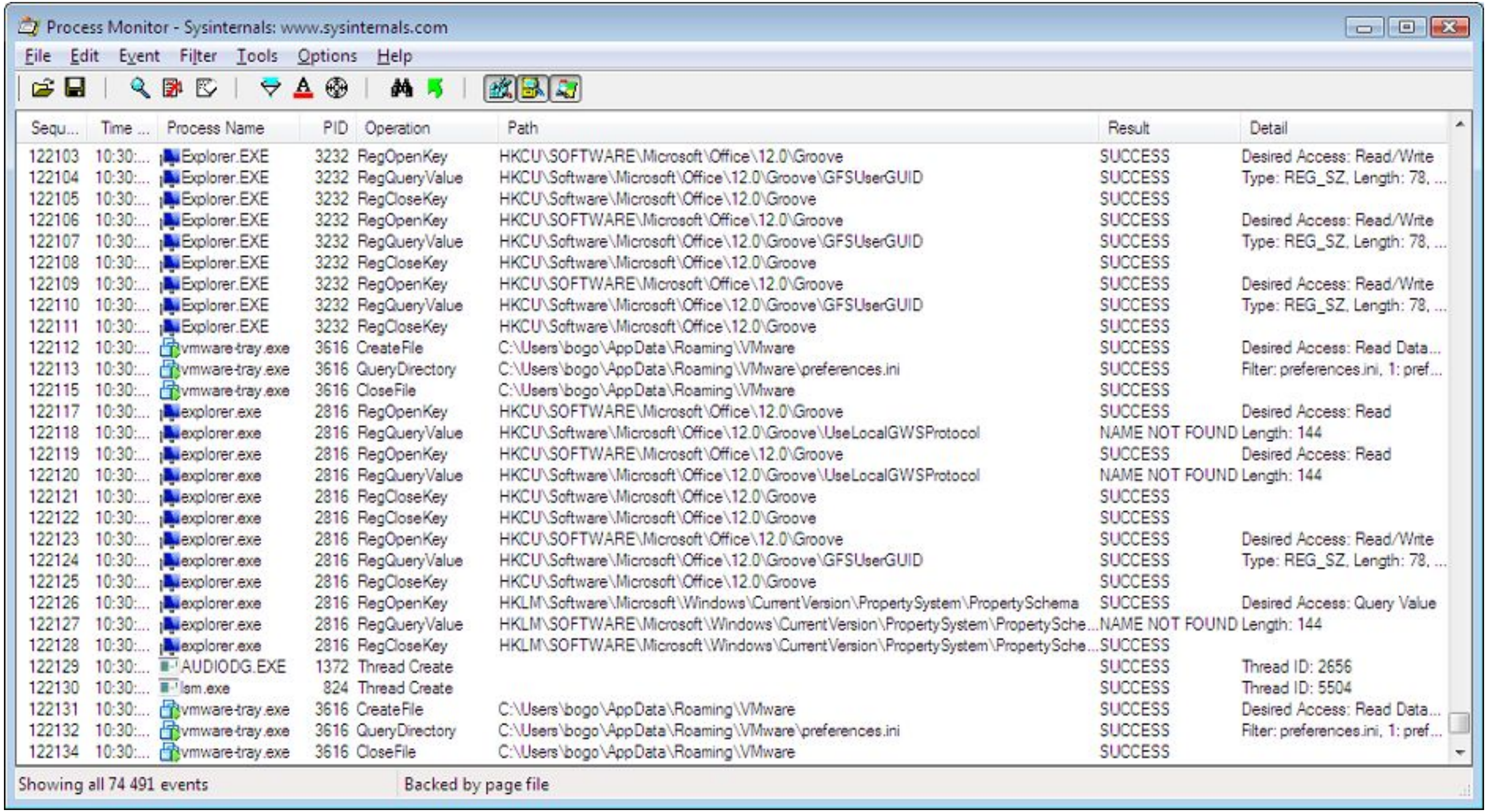
Motivation

- Issues with manual analysis
  - Require an experienced analyst
  - Time consuming and tedious

- **Project Goal:** Automate the process of software analysis with the use of data analytics on logs to point analysts to interesting behavior.
  - Begin by researching the viability of different analysis methods.
  - Could these methods identify what behaviors the program exhibited given only the logs?

Note: All data in this presentation is synthetic, but is representative of the real data and results.

The Data



Sequ...	Time ...	Process Name	PID	Operation	Path	Result	Detail
122103	10:30:...	Explorer.EXE	3232	RegOpenKey	HKCU\SOFTWARE\Microsoft\Office\12.0\Groove	SUCCESS	Desired Access: Read/Write
122104	10:30:...	Explorer.EXE	3232	RegQueryValue	HKCU\Software\Microsoft\Office\12.0\Groove\GFSUserGUID	SUCCESS	Type: REG_SZ, Length: 78, ...
122105	10:30:...	Explorer.EXE	3232	RegCloseKey	HKCU\Software\Microsoft\Office\12.0\Groove	SUCCESS	
122106	10:30:...	Explorer.EXE	3232	RegOpenKey	HKCU\SOFTWARE\Microsoft\Office\12.0\Groove	SUCCESS	Desired Access: Read/Write
122107	10:30:...	Explorer.EXE	3232	RegQueryValue	HKCU\Software\Microsoft\Office\12.0\Groove\GFSUserGUID	SUCCESS	Type: REG_SZ, Length: 78, ...
122108	10:30:...	Explorer.EXE	3232	RegCloseKey	HKCU\Software\Microsoft\Office\12.0\Groove	SUCCESS	
122109	10:30:...	Explorer.EXE	3232	RegOpenKey	HKCU\SOFTWARE\Microsoft\Office\12.0\Groove	SUCCESS	Desired Access: Read/Write
122110	10:30:...	Explorer.EXE	3232	RegQueryValue	HKCU\Software\Microsoft\Office\12.0\Groove\GFSUserGUID	SUCCESS	Type: REG_SZ, Length: 78, ...
122111	10:30:...	Explorer.EXE	3232	RegCloseKey	HKCU\Software\Microsoft\Office\12.0\Groove	SUCCESS	
122112	10:30:...	vmware-tray.exe	3616	CreateFile	C:\Users\bogo\AppData\Roaming\VMware	SUCCESS	Desired Access: Read Data...
122113	10:30:...	vmware-tray.exe	3616	QueryDirectory	C:\Users\bogo\AppData\Roaming\VMware\preferences.ini	SUCCESS	Filter: preferences.ini, 1: pref...
122115	10:30:...	vmware-tray.exe	3616	CloseFile	C:\Users\bogo\AppData\Roaming\VMware	SUCCESS	
122117	10:30:...	explorer.exe	2816	RegOpenKey	HKCU\SOFTWARE\Microsoft\Office\12.0\Groove	SUCCESS	Desired Access: Read
122118	10:30:...	explorer.exe	2816	RegQueryValue	HKCU\Software\Microsoft\Office\12.0\Groove\UseLocalGWSProtocol	NAME NOT FOUND	Length: 144
122119	10:30:...	explorer.exe	2816	RegOpenKey	HKCU\SOFTWARE\Microsoft\Office\12.0\Groove	SUCCESS	Desired Access: Read
122120	10:30:...	explorer.exe	2816	RegQueryValue	HKCU\Software\Microsoft\Office\12.0\Groove\UseLocalGWSProtocol	NAME NOT FOUND	Length: 144
122121	10:30:...	explorer.exe	2816	RegCloseKey	HKCU\Software\Microsoft\Office\12.0\Groove	SUCCESS	
122122	10:30:...	explorer.exe	2816	RegCloseKey	HKCU\Software\Microsoft\Office\12.0\Groove	SUCCESS	
122123	10:30:...	explorer.exe	2816	RegOpenKey	HKCU\SOFTWARE\Microsoft\Office\12.0\Groove	SUCCESS	Desired Access: Read/Write
122124	10:30:...	explorer.exe	2816	RegQueryValue	HKCU\Software\Microsoft\Office\12.0\Groove\GFSUserGUID	SUCCESS	Type: REG_SZ, Length: 78, ...
122125	10:30:...	explorer.exe	2816	RegCloseKey	HKCU\Software\Microsoft\Office\12.0\Groove	SUCCESS	
122126	10:30:...	explorer.exe	2816	RegOpenKey	HKLM\Software\Microsoft\Windows\CurrentVersion\PropertySystem\PropertySchema	SUCCESS	Desired Access: Query Value
122127	10:30:...	explorer.exe	2816	RegQueryValue	HKLM\SOFTWARE\Microsoft\Windows\CurrentVersion\PropertySystem\PropertySche...	NAME NOT FOUND	Length: 144
122128	10:30:...	explorer.exe	2816	RegCloseKey	HKLM\SOFTWARE\Microsoft\Windows\CurrentVersion\PropertySystem\PropertySche...	SUCCESS	
122129	10:30:...	AUDIODG.EXE	1372	Thread Create		SUCCESS	Thread ID: 2656
122130	10:30:...	lsm.exe	824	Thread Create		SUCCESS	Thread ID: 5504
122131	10:30:...	vmware-tray.exe	3616	CreateFile	C:\Users\bogo\AppData\Roaming\VMware	SUCCESS	Desired Access: Read Data...
122132	10:30:...	vmware-tray.exe	3616	QueryDirectory	C:\Users\bogo\AppData\Roaming\VMware\preferences.ini	SUCCESS	Filter: preferences.ini, 1: pref...
122134	10:30:...	vmware-tray.exe	3616	CloseFile	C:\Users\bogo\AppData\Roaming\VMware	SUCCESS	

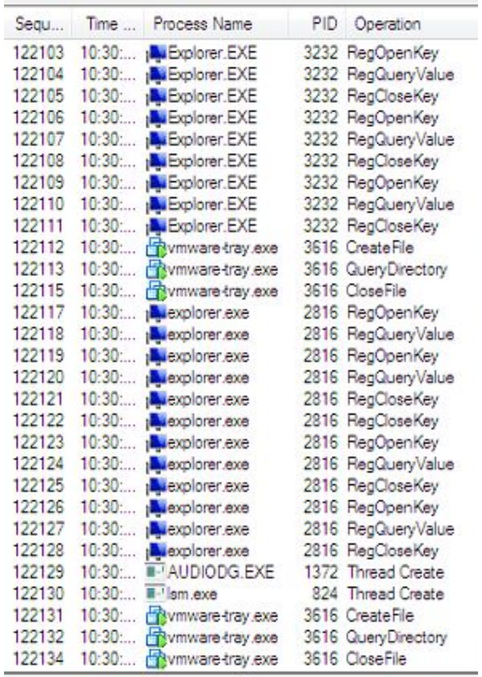
Showing all 74 491 events

Backed by page file

Motivation for Natural Language Processing

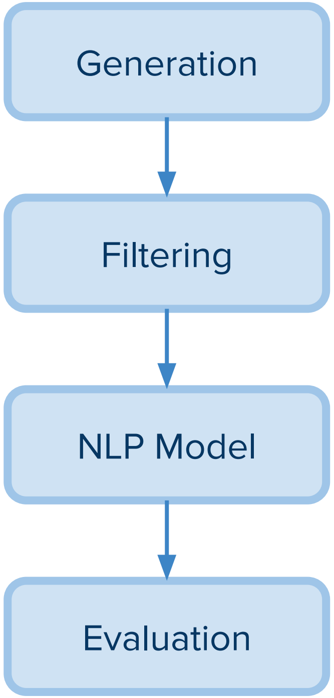
- Can think of logs like a collection of text
  - Each event (or row) in the log can be thought of as a word

- **Hypothesis:** Seeing certain words in a particular sequence, or even having certain words just being present in the log, could tell us something about the behavior of the program that generated that log



```
Sequ...  Time ...  Process Name      PID   Operation
122103   10:30:... Explorer.EXE      3232  RegOpenKey
122104   10:30:... Explorer.EXE      3232  RegQueryValue
122105   10:30:... Explorer.EXE      3232  RegCloseKey
122106   10:30:... Explorer.EXE      3232  RegOpenKey
122107   10:30:... Explorer.EXE      3232  RegQueryValue
122108   10:30:... Explorer.EXE      3232  RegCloseKey
122109   10:30:... Explorer.EXE      3232  RegOpenKey
122110   10:30:... Explorer.EXE      3232  RegQueryValue
122111   10:30:... Explorer.EXE      3232  RegCloseKey
122112   10:30:... vmware-tray.exe   3616  CreateFile
122113   10:30:... vmware-tray.exe   3616  QueryDirectory
122115   10:30:... vmware-tray.exe   3616  CloseFile
122117   10:30:... explorer.exe      2816  RegOpenKey
122118   10:30:... explorer.exe      2816  RegQueryValue
122119   10:30:... explorer.exe      2816  RegOpenKey
122120   10:30:... explorer.exe      2816  RegQueryValue
122121   10:30:... explorer.exe      2816  RegCloseKey
122122   10:30:... explorer.exe      2816  RegCloseKey
122123   10:30:... explorer.exe      2816  RegOpenKey
122124   10:30:... explorer.exe      2816  RegQueryValue
122125   10:30:... explorer.exe      2816  RegCloseKey
122126   10:30:... explorer.exe      2816  RegOpenKey
122127   10:30:... explorer.exe      2816  RegQueryValue
122128   10:30:... explorer.exe      2816  RegCloseKey
122129   10:30:... AUDIODG.EXE       1372  Thread Create
122130   10:30:... lsm.exe            824  Thread Create
122131   10:30:... vmware-tray.exe   3616  CreateFile
122132   10:30:... vmware-tray.exe   3616  QueryDirectory
122134   10:30:... vmware-tray.exe   3616  CloseFile
```

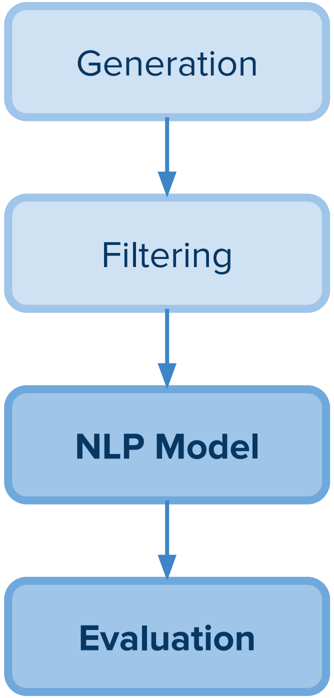
Experimental Design



In order to research ways to automate the analysis, we used the following pipeline:

- **Generate** logs for different types of program behavior
  - We focused on two programs:
    - Notepad
    - Windows Defender (Online scan VS offline scan)
- **Filter** Logs
- Feed logs into **NLP Model**
- **Evaluate** results
  - Are we able to differentiate different types of behavior based on the model?

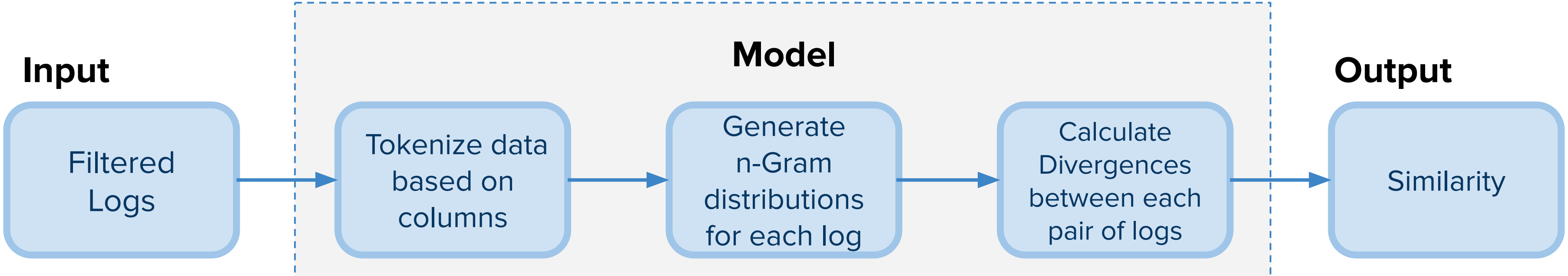
Experimental Design



In order to research ways to automate the analysis, we used the following pipeline:

- **Generate** logs for different types of program behavior
  - We focused on two programs:
    - Notepad
    - Windows Defender (Online scan VS offline scan)
- **Filter** Logs
- Feed logs into **NLP Model**
- **Evaluate** results
  - Are we able to differentiate different types of behavior based on the model?

Model 1: Pairwise n-gram divergence comparisons



Input

Filtered Logs

Model

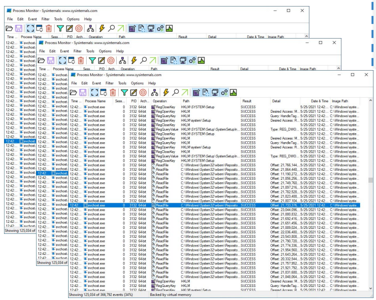
Tokenize data based on columns

Generate n-Gram distributions for each log

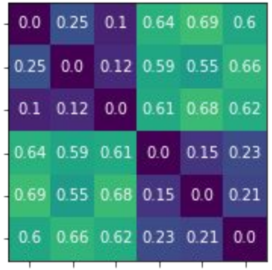
Calculate Divergences between each pair of logs

Output

Similarity

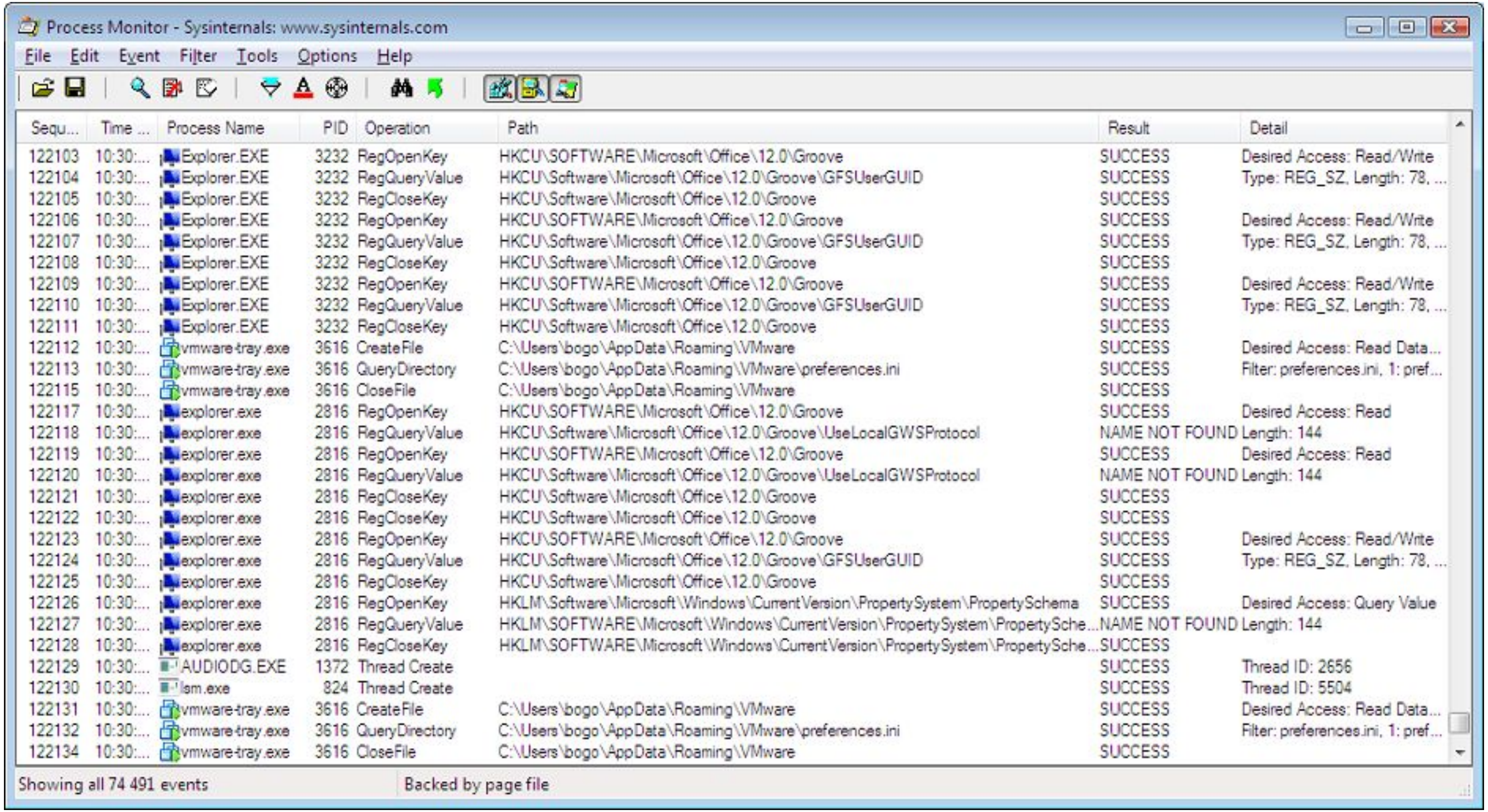


```
Time ...  Process Name  Sess...  PID   Arch...  Operation      Path                                  Result   Detail               Date & Time          Image Path
12:42:... svchost.exe   0        3132  64-bit   RegCloseKey    HKLM\SYSTEM\Setup                     SUCCESS                       5/25/2021 12:42:...  C:\Windows\syste...
12:42:... svchost.exe   0        3132  64-bit   RegOpenKey     HKLM                                  SUCCESS  Desired Access: M... 5/25/2021 12:42:...  C:\Windows\syste...
12:42:... svchost.exe   0        3132  64-bit   RegQueryKey    HKLM                                  SUCCESS  Query: HandleTag...  5/25/2021 12:42:...  C:\Windows\syste...
12:42:... svchost.exe   0        3132  64-bit   RegOpenKey     HKLM\system\Setup                     SUCCESS  Desired Access: R... 5/25/2021 12:42:...  C:\Windows\syste...
12:42:... svchost.exe   0        3132  64-bit   RegCloseKey    HKLM                                  SUCCESS                       5/25/2021 12:42:...  C:\Windows\syste...
12:42:... svchost.exe   0        3132  64-bit   RegQueryValue  HKLM\SYSTEM\Setup\SystemSetupIn...    SUCCESS  Type: REG_DWO...     5/25/2021 12:42:...  C:\Windows\syste...
12:42:... svchost.exe   0        3132  64-bit   RegCloseKey    HKLM\SYSTEM\Setup                     SUCCESS                       5/25/2021 12:42:...  C:\Windows\syste...
12:42:... svchost.exe   0        3132  64-bit   RegOpenKey     HKLM                                  SUCCESS  Desired Access: M... 5/25/2021 12:42:...  C:\Windows\syste...
12:42:... svchost.exe   0        3132  64-bit   RegQueryKey    HKLM                                  SUCCESS  Query: HandleTag...  5/25/2021 12:42:...  C:\Windows\syste...
12:42:... svchost.exe   0        3132  64-bit   RegOpenKey     HKLM\system\Setup                     SUCCESS  Desired Access: R... 5/25/2021 12:42:...  C:\Windows\syste...
12:42:... svchost.exe   0        3132  64-bit   RegCloseKey    HKLM                                  SUCCESS                       5/25/2021 12:42:...  C:\Windows\syste...
12:42:... svchost.exe   0        3132  64-bit   RegQueryValue  HKLM\SYSTEM\Setup\SystemSetupIn...    SUCCESS  Type: REG_DWO...     5/25/2021 12:42:...  C:\Windows\syste...
12:42:... svchost.exe   0        3132  64-bit   RegCloseKey    HKLM\SYSTEM\Setup                     SUCCESS                       5/25/2021 12:42:...  C:\Windows\syste...
12:42:... svchost.exe   0        3132  64-bit   ReadFile       C:\Windows\System32\wbem\Reposito...  SUCCESS  Offset: 21,766,144,...5/25/2021 12:42:...  C:\Windows\syste...
12:42:... svchost.exe   0        3132  64-bit   ReadFile       C:\Windows\System32\wbem\Reposito...  SUCCESS  Offset: 21,864,448,...5/25/2021 12:42:...  C:\Windows\syste...
12:42:... svchost.exe   0        3132  64-bit   ReadFile       C:\Windows\System32\wbem\Reposito...  SUCCESS  Offset: 11,190,272,...5/25/2021 12:42:...  C:\Windows\syste...
12:42:... svchost.exe   0        3132  64-bit   ReadFile       C:\Windows\System32\wbem\Reposito...  SUCCESS  Offset: 21,856,256,...5/25/2021 12:42:...  C:\Windows\syste...
12:42:... svchost.exe   0        3132  64-bit   ReadFile       C:\Windows\System32\wbem\Reposito...  SUCCESS  Offset: 21,749,760,...5/25/2021 12:42:...  C:\Windows\syste...
12:42:... svchost.exe   0        3132  64-bit   ReadFile       C:\Windows\System32\wbem\Reposito...  SUCCESS  Offset: 21,897,216,...5/25/2021 12:42:...  C:\Windows\syste...
12:42:... svchost.exe   0        3132  64-bit   ReadFile       C:\Windows\System32\wbem\Reposito...  SUCCESS  Offset: 21,782,528,...5/25/2021 12:42:...  C:\Windows\syste...
12:42:... svchost.exe   0        3132  64-bit   ReadFile       C:\Windows\System32\wbem\Reposito...  SUCCESS  Offset: 21,823,488,...5/25/2021 12:42:...  C:\Windows\syste...
12:42:... svchost.exe   0        3132  64-bit   ReadFile       C:\Windows\System32\wbem\Reposito...  SUCCESS  Offset: 21,807,104,...5/25/2021 12:42:...  C:\Windows\syste...
12:42:... svchost.exe   0        3132  64-bit   ReadFile       C:\Windows\System32\wbem\Reposito...  SUCCESS  Offset: 21,733,376,...5/25/2021 12:42:...  C:\Windows\syste...
12:42:... svchost.exe   0        3132  64-bit   ReadFile       C:\Windows\System32\wbem\Reposito...  SUCCESS  Offset: 23,044,096,...5/25/2021 12:42:...  C:\Windows\syste...
12:42:... svchost.exe   0        3132  64-bit   ReadFile       C:\Windows\System32\wbem\Reposito...  SUCCESS  Offset: 21,880,832,...5/25/2021 12:42:...  C:\Windows\syste...
12:42:... svchost.exe   0        3132  64-bit   ReadFile       C:\Windows\System32\wbem\Reposito...  SUCCESS  Offset: 21,692,416,...5/25/2021 12:42:...  C:\Windows\syste...
12:42:... svchost.exe   0        3132  64-bit   ReadFile       C:\Windows\System32\wbem\Reposito...  SUCCESS  Offset: 21,651,456,...5/25/2021 12:42:...  C:\Windows\syste...
12:42:... svchost.exe   0        3132  64-bit   ReadFile       C:\Windows\System32\wbem\Reposito...  SUCCESS  Offset: 21,889,024,...5/25/2021 12:42:...  C:\Windows\syste...
12:42:... svchost.exe   0        3132  64-bit   ReadFile       C:\Windows\System32\wbem\Reposito...  SUCCESS  Offset: 22,036,480,...5/25/2021 12:42:...  C:\Windows\syste...
12:42:... svchost.exe   0        3132  64-bit   ReadFile       C:\Windows\System32\wbem\Reposito...  SUCCESS  Offset: 23,543,808,...5/25/2021 12:42:...  C:\Windows\syste...
12:42:... svchost.exe   0        3132  64-bit   ReadFile       C:\Windows\System32\wbem\Reposito...  SUCCESS  Offset: 21,790,720,...5/25/2021 12:42:...  C:\Windows\syste...
12:42:... svchost.exe   0        3132  64-bit   ReadFile       C:\Windows\System32\wbem\Reposito...  SUCCESS  Offset: 21,774,336,...5/25/2021 12:42:...  C:\Windows\syste...
12:42:... svchost.exe   0        3132  64-bit   ReadFile       C:\Windows\System32\wbem\Reposito...  SUCCESS  Offset: 21,954,560,...5/25/2021 12:42:...  C:\Windows\syste...
12:42:... svchost.exe   0        3132  64-bit   ReadFile       C:\Windows\System32\wbem\Reposito...  SUCCESS  Offset: 21,643,264,...5/25/2021 12:42:...  C:\Windows\syste...
12:42:... svchost.exe   0        3132  64-bit   ReadFile       C:\Windows\System32\wbem\Reposito...  SUCCESS  Offset: 20,332,544,...5/25/2021 12:42:...  C:\Windows\syste...
12:42:... svchost.exe   0        3132  64-bit   ReadFile       C:\Windows\System32\wbem\Reposito...  SUCCESS  Offset: 21,757,952,...5/25/2021 12:42:...  C:\Windows\syste...
12:42:... svchost.exe   0        3132  64-bit   ReadFile       C:\Windows\System32\wbem\Reposito...  SUCCESS  Offset: 21,921,792,...5/25/2021 12:42:...  C:\Windows\syste...
12:42:... svchost.exe   0        3132  64-bit   ReadFile       C:\Windows\System32\wbem\Reposito...  SUCCESS  Offset: 21,831,680,...5/25/2021 12:42:...  C:\Windows\syste...
12:42:... svchost.exe   0        3132  64-bit   ReadFile       C:\Windows\System32\wbem\Reposito...  SUCCESS  Offset: 21,848,064,...5/25/2021 12:42:...  C:\Windows\syste...
12:42:... svchost.exe   0        3132  64-bit   RegOpenKey     HKLM                                  SUCCESS  Desired Access: M... 5/25/2021 12:42:...  C:\Windows\syste...
12:42:... svchost.exe   0        3132  64-bit   RegQueryKey    HKLM                                  SUCCESS  Query: HandleTag...  5/25/2021 12:42:...  C:\Windows\syste...
12:42:... svchost.exe   0        3132  64-bit   RegOpenKey     HKLM\system\Setup                     SUCCESS  Desired Access: R... 5/25/2021 12:42:...  C:\Windows\syste...
Showing 125,034 of 366,792 events (34%)   Backed by virtual memory
```



0.0	0.25	0.1	0.64	0.69	0.6
0.25	0.0	0.12	0.59	0.55	0.66
0.1	0.12	0.0	0.61	0.68	0.62
0.64	0.59	0.61	0.0	0.15	0.23
0.69	0.55	0.68	0.15	0.0	0.21
0.6	0.66	0.62	0.23	0.21	0.0

The Data



Sequ...	Time ...	Process Name	PID	Operation	Path	Result	Detail
122103	10:30:...	Explorer.EXE	3232	RegOpenKey	HKCU\SOFTWARE\Microsoft\Office\12.0\Groove	SUCCESS	Desired Access: Read/Write
122104	10:30:...	Explorer.EXE	3232	RegQueryValue	HKCU\Software\Microsoft\Office\12.0\Groove\GFSUserGUID	SUCCESS	Type: REG_SZ, Length: 78, ...
122105	10:30:...	Explorer.EXE	3232	RegCloseKey	HKCU\Software\Microsoft\Office\12.0\Groove	SUCCESS	
122106	10:30:...	Explorer.EXE	3232	RegOpenKey	HKCU\SOFTWARE\Microsoft\Office\12.0\Groove	SUCCESS	Desired Access: Read/Write
122107	10:30:...	Explorer.EXE	3232	RegQueryValue	HKCU\Software\Microsoft\Office\12.0\Groove\GFSUserGUID	SUCCESS	Type: REG_SZ, Length: 78, ...
122108	10:30:...	Explorer.EXE	3232	RegCloseKey	HKCU\Software\Microsoft\Office\12.0\Groove	SUCCESS	
122109	10:30:...	Explorer.EXE	3232	RegOpenKey	HKCU\SOFTWARE\Microsoft\Office\12.0\Groove	SUCCESS	Desired Access: Read/Write
122110	10:30:...	Explorer.EXE	3232	RegQueryValue	HKCU\Software\Microsoft\Office\12.0\Groove\GFSUserGUID	SUCCESS	Type: REG_SZ, Length: 78, ...
122111	10:30:...	Explorer.EXE	3232	RegCloseKey	HKCU\Software\Microsoft\Office\12.0\Groove	SUCCESS	
122112	10:30:...	vmware-tray.exe	3616	CreateFile	C:\Users\bogo\AppData\Roaming\VMware	SUCCESS	Desired Access: Read Data...
122113	10:30:...	vmware-tray.exe	3616	QueryDirectory	C:\Users\bogo\AppData\Roaming\VMware\preferences.ini	SUCCESS	Filter: preferences.ini, 1: pref...
122115	10:30:...	vmware-tray.exe	3616	CloseFile	C:\Users\bogo\AppData\Roaming\VMware	SUCCESS	
122117	10:30:...	explorer.exe	2816	RegOpenKey	HKCU\SOFTWARE\Microsoft\Office\12.0\Groove	SUCCESS	Desired Access: Read
122118	10:30:...	explorer.exe	2816	RegQueryValue	HKCU\Software\Microsoft\Office\12.0\Groove\UseLocalGWSProtocol	NAME NOT FOUND	Length: 144
122119	10:30:...	explorer.exe	2816	RegOpenKey	HKCU\SOFTWARE\Microsoft\Office\12.0\Groove	SUCCESS	Desired Access: Read
122120	10:30:...	explorer.exe	2816	RegQueryValue	HKCU\Software\Microsoft\Office\12.0\Groove\UseLocalGWSProtocol	NAME NOT FOUND	Length: 144
122121	10:30:...	explorer.exe	2816	RegCloseKey	HKCU\Software\Microsoft\Office\12.0\Groove	SUCCESS	
122122	10:30:...	explorer.exe	2816	RegCloseKey	HKCU\Software\Microsoft\Office\12.0\Groove	SUCCESS	
122123	10:30:...	explorer.exe	2816	RegOpenKey	HKCU\SOFTWARE\Microsoft\Office\12.0\Groove	SUCCESS	Desired Access: Read/Write
122124	10:30:...	explorer.exe	2816	RegQueryValue	HKCU\Software\Microsoft\Office\12.0\Groove\GFSUserGUID	SUCCESS	Type: REG_SZ, Length: 78, ...
122125	10:30:...	explorer.exe	2816	RegCloseKey	HKCU\Software\Microsoft\Office\12.0\Groove	SUCCESS	
122126	10:30:...	explorer.exe	2816	RegOpenKey	HKLM\Software\Microsoft\Windows\CurrentVersion\PropertySystem\PropertySchema	SUCCESS	Desired Access: Query Value
122127	10:30:...	explorer.exe	2816	RegQueryValue	HKLM\SOFTWARE\Microsoft\Windows\CurrentVersion\PropertySystem\PropertySche...	NAME NOT FOUND	Length: 144
122128	10:30:...	explorer.exe	2816	RegCloseKey	HKLM\SOFTWARE\Microsoft\Windows\CurrentVersion\PropertySystem\PropertySche...	SUCCESS	
122129	10:30:...	AUDIODG.EXE	1372	Thread Create		SUCCESS	Thread ID: 2656
122130	10:30:...	lsm.exe	824	Thread Create		SUCCESS	Thread ID: 5504
122131	10:30:...	vmware-tray.exe	3616	CreateFile	C:\Users\bogo\AppData\Roaming\VMware	SUCCESS	Desired Access: Read Data...
122132	10:30:...	vmware-tray.exe	3616	QueryDirectory	C:\Users\bogo\AppData\Roaming\VMware\preferences.ini	SUCCESS	Filter: preferences.ini, 1: pref...
122134	10:30:...	vmware-tray.exe	3616	CloseFile	C:\Users\bogo\AppData\Roaming\VMware	SUCCESS	

Showing all 74 491 events

Backed by page file

Tokenizing Logs

- We’re using NLP, but what is a “word” considered in a log?
- Say we are given the super short example log below
- All columns might not be relevant to the analysis -- chose only relevant ones (this was something we experimented with)

Sequence	Time	Process Name	PID	Operation	Path	Result	Detail
122103	10:30…	Explorer.EXE	3232	RegOpenKey	C:\Users\…	SUCCESS	Desired Access: Read/Write
122104	10:30…	Explorer.EXE	3232	RegQueryValue	C:\Users\…	SUCCESS	Type: REG_SZ…

Tokenizing Logs

- We’re using NLP, but what is a “word” considered in a log?
- Say we are given the super short example log below
- All columns might not be relevant to the analysis -- chose only relevant ones (this was something we experimented with)

Sequence	Time	Process Name	PID	Operation	Path	Result	Detail
122103	10:30…	Explorer.EXE	3232	RegOpenKey	C:\Users\...	SUCCESS	Desired Access: Read/Write
122104	10:30…	Explorer.EXE	3232	RegQueryValue	C:\Users\...	SUCCESS	Type: REG_SZ…

Tokenizing Logs

- We’re using NLP, but what is a “word” considered in a log?
- Say we are given the super short example log below
- All columns might not be relevant to the analysis -- chose only relevant ones (this was something we experimented with)

Sequence	Time	Process Name	PID	Operation	Path	Result	Detail
122103	10:30…	Explorer.EXE	3232	RegOpenKey	C:\Users\…	SUCCESS	Desired Access: Read/Write
122104	10:30…	Explorer.EXE	3232	RegQueryValue	C:\Users\…	SUCCESS	Type: REG_SZ…

[(Explorer.EXE, RegOpenKey, SUCCESS), (Explorer.EXE, RegQueryValue, SUCCESS)]

What is an n-gram?

- A n-gram is a contiguous sequence of n items from a given text.
- For example the 2-grams of this sequence:

“sphinx of black quartz judge my vow”

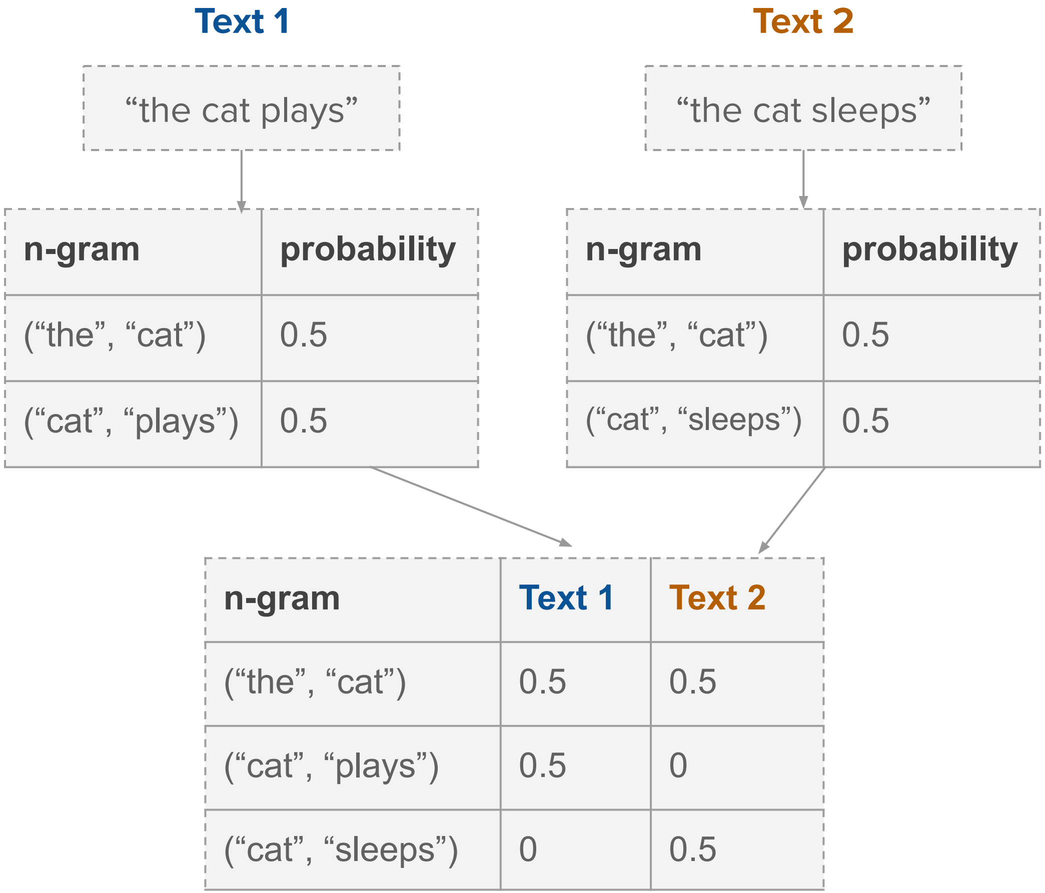
- Would be:

(“sphinx”, “of”), (“of”, “black”), (“black”, “quartz”), (“quartz”, “judge”), (“judge”, “my”), (“my, vow”)

- Why n-grams?
  - n-grams can capture sequences of words

n-gram distributions

- Say we have two texts we’d like to compare
- We generate the n-grams
  - 2-grams in this case
- Generate n-gram distributions
- And compare the distributions as vectors



Text 1

“the cat plays”

n-gram	probability
(“the”, “cat”)	0.5
(“cat”, “plays”)	0.5

Text 2

“the cat sleeps”

n-gram	probability
(“the”, “cat”)	0.5
(“cat”, “sleeps”)	0.5

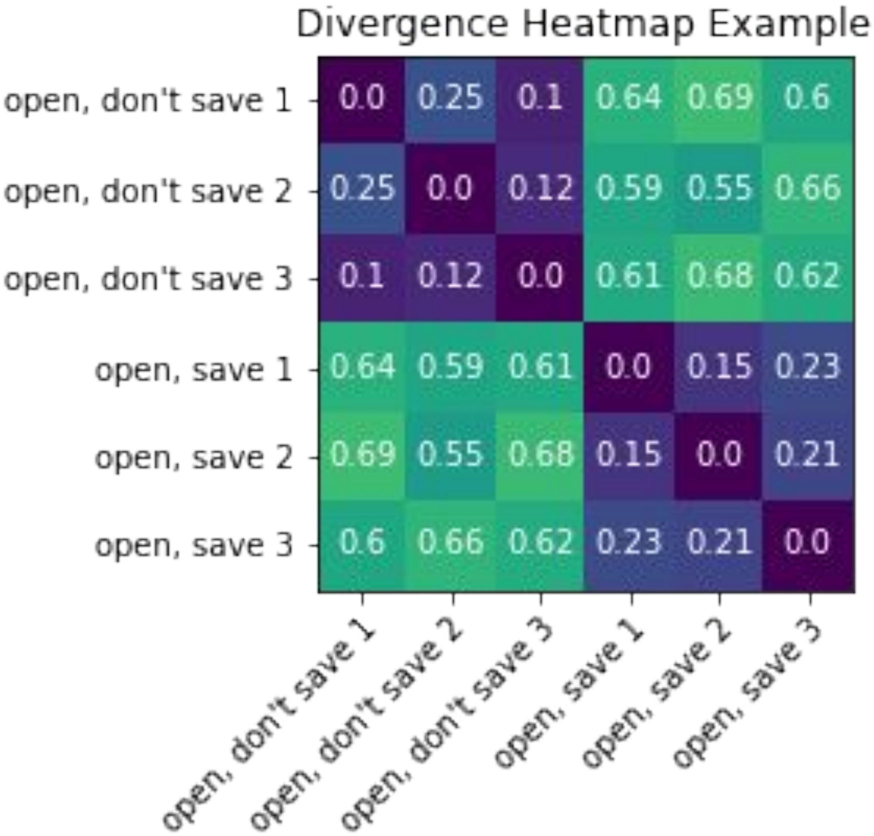
n-gram	Text 1	Text 2
(“the”, “cat”)	0.5	0.5
(“cat”, “plays”)	0.5	0
(“cat”, “sleeps”)	0	0.5

Divergences

- A divergence function is a function which calculates the “distance” of one probability distribution to another
- Gives us a numerical way to compare texts by comparing n-gram distributions
- Divergences Used
  - Bhattacharyya distance
    - Results can be between 0 and infinity
  - Jensen-Shannon Divergence
    - Results can be between 0 and 1. (We used this distance for this reason)

Heatmap Visualization

- Example of what a heatmap might look like
  - Comparison of 2 cases with 3 replicates each
  - For a simple program like notepad



Divergence Heatmap Example

	open, don't save 1	open, don't save 2	open, don't save 3	open, save 1	open, save 2	open, save 3
open, don't save 1	0.0	0.25	0.1	0.64	0.69	0.6
open, don't save 2	0.25	0.0	0.12	0.59	0.55	0.66
open, don't save 3	0.1	0.12	0.0	0.61	0.68	0.62
open, save 1	0.64	0.59	0.61	0.0	0.15	0.23
open, save 2	0.69	0.55	0.68	0.15	0.0	0.21
open, save 3	0.6	0.66	0.62	0.23	0.21	0.0

Additional Techniques

- **n-gram explainability**
  - Output a file which will order the n-grams by what contributed most
  - Could help analysts understand what exactly made cases different from each other

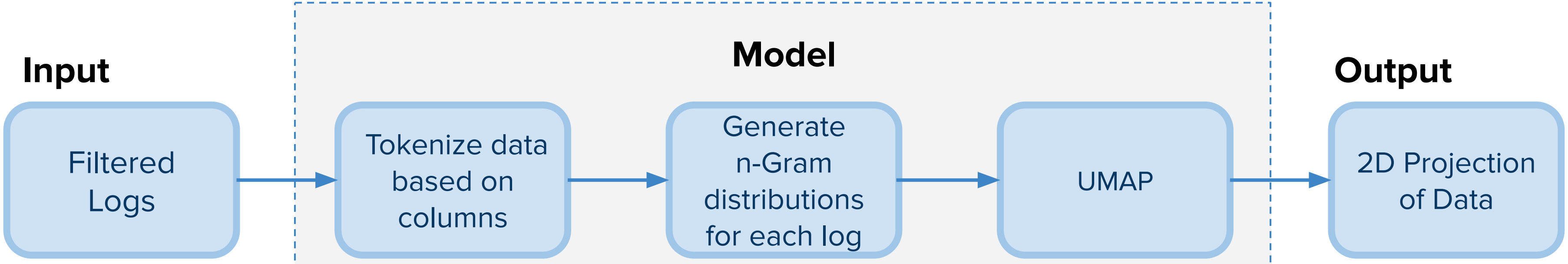
Additional Techniques

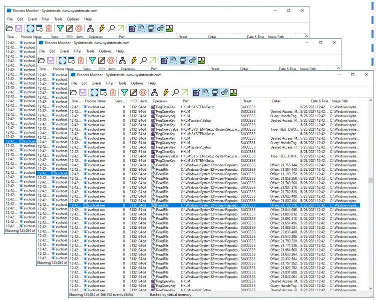
- **k-fold comparisons**
  - A way to compare cases and see if they are distinguishable
  - Algorithm:
    - For each dataset of cases generated from the same behavior:
      - Split the dataset into k groups
      - For each unique group
        - Take the group and separate it from the others. Call it item 1
        - Call the remaining groups item2
        - Find the divergence between item1 and item2
    - Take the maximum divergence seen.
  - This is the maximum allowable divergence between cases that are the same. It follows that when comparing two cases, if the divergence is greater than that number, then the two cases are different.

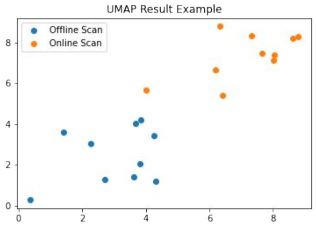
Results of Model 1

- Works great for simple cases
  - ie Notepad
- Breaks down for more complex cases
  - Cannot distinguish
  - For example a defender online vs offline scan

Model 2: dimensionality reduction with UMAP


Input
Filtered Logs
Model
Tokenize data based on columns
Generate n-Gram distributions for each log
UMAP
Output
2D Projection of Data


Process Monitor - Sysinternals: www.sysinternals.com
File Edit Event Filter Tools Options Help



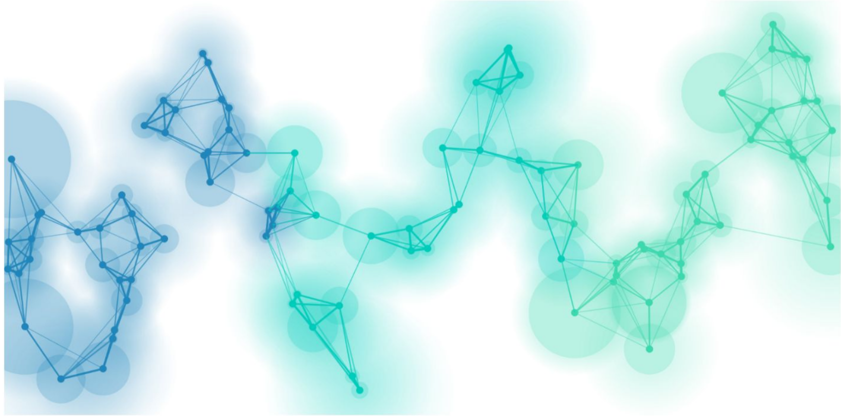
UMAP

- **U**niform **Ma**nifold **A**pproximation and **P**rojection
- Dimensionality reduction algorithm (like PCA or tSNE)
- Two steps:
  - Construct a high dimensional graph representation of the data
  - Optimize a low-dimensional graph to be as structurally similar as possible
- In our project, we were reducing the dimensions of the n-gram distribution vectors
  - “points” refers to the n-gram distribution vectors

UMAP

Step 1: Construct a high dimensional graph representation of the data by building a "fuzzy simplicial complex"

- What is a "fuzzy simplicial complex"?
  - Weighted graph, with edge weights representing the likelihood that two points are connected
- How does it build this?
  - Each point has a radius extend from it.
  - Two points are considered connected when those radii overlap.
  - The radius size is based on the distance to each point's nth nearest neighbor
  - Graph is "fuzzy" because the likelihood of connection decreases as the radius grows
- Demo (by Andy Coenen, Adam Pearce)

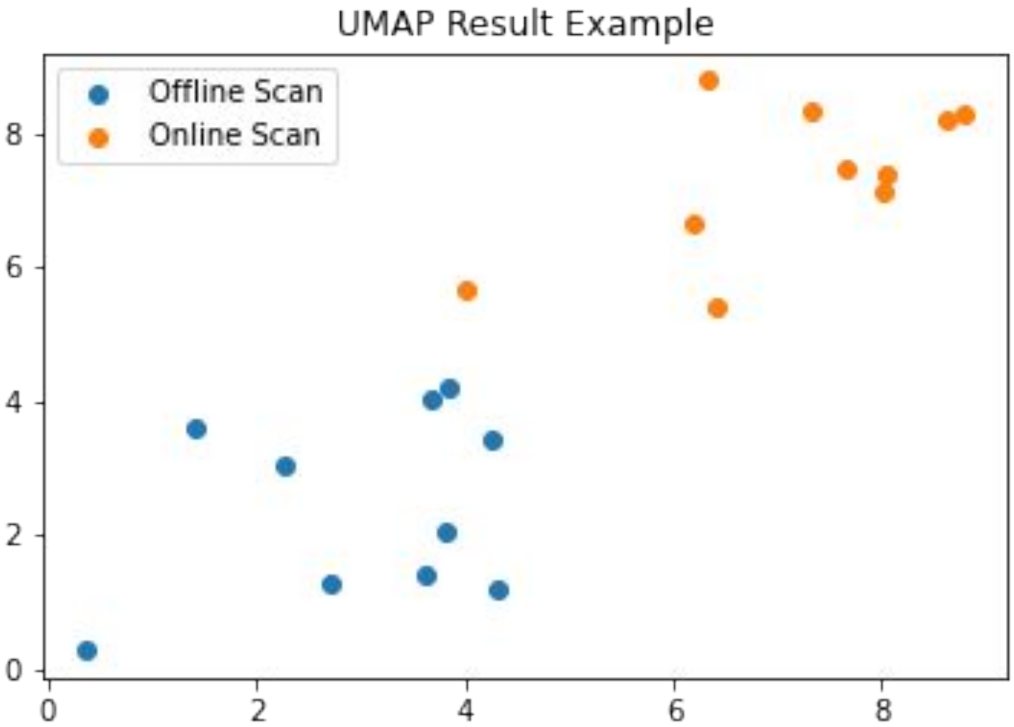


UMAP

Step 2: Optimize a low-dimensional graph to be as structurally similar as possible

- Imagine the high dimensional graph as if the edges between points were springs, where each spring is stronger as the edge probability increases
- Then we squish it down into smaller dimensions

UMAP Visualization



Future Directions

- Test UMAP with different behavior cases
- Try UMAP with different hyperparameters
    - n_neighbors, min_dist
- Look into what information could be gained by looking into the stack traces
    - We were looking at flat csvs
    - Process manager captures the stack trace for each event
- Attempt to classify cases based on UMAP results
    - Clustering, logistic regression, neural networks, etc.

Questions

joelynelson3333@gmail.com