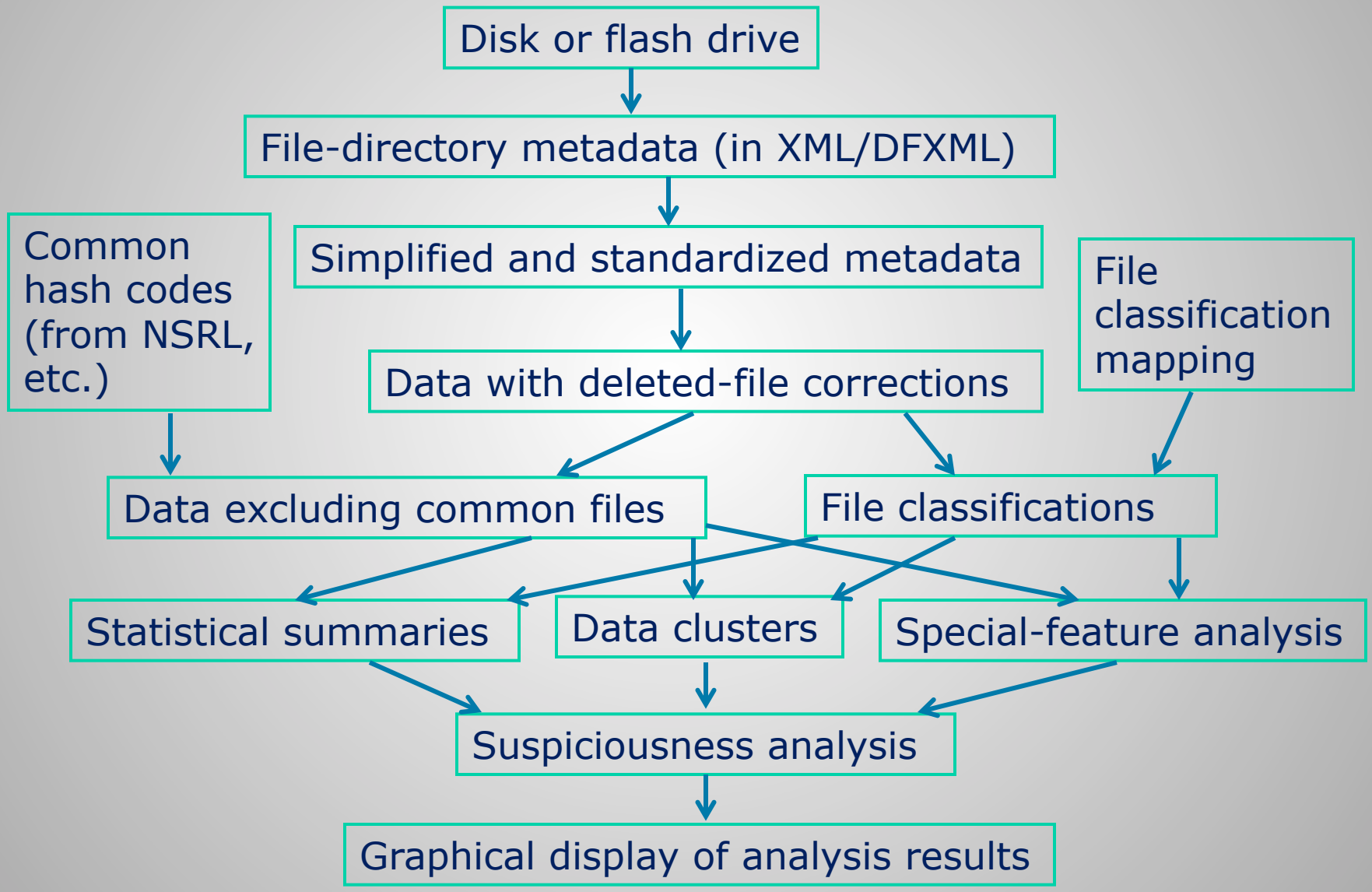
Testing the National Software Reference Library

Neil C. Rowe
U.S. Naval Postgraduate School
Monterey, California, USA
ncrowe@nps.edu

Forensics of directory metadata

- We need tools to quickly find key information on a drive without searching file contents.
- File and directory metadata is a big help to characterize drives (or partitions on a cloud).
- We are developing a tool "Dirim".
- Our testbed the "Real Drive Corpus" is purchased from 22 countries, mostly China, Mexico, Israel, Palestine, and India – now 2420 drives and 44 million files.
- It also includes wireless and storage devices.
- For analysis, we exclude files with hashes found in the National Software Reference Library Reference Data Set (NSRL RDS) – it removes 30% of the files – and 5% of the hashes.
- Research question: Just how good is the NSRL?

The Dirim file-metadata analysis system



File metadata we extract from a disk image

Ordinal features	Nominal features	Boolean features
File size	Drive name	Allocated?
Access minus creation time	File name	Compressed?
Access minus modification time	File extension	Encrypted?
Modification minus creation time	Top-level directory	Empty?
Depth in file hierarchy	Immediate directory	Much punctuation?
Number of fragments	Hash code	Many digits?
Size of directory	Product classification (from NSRL data)	Unicode characters?
Frequency of file in the corpus		Punctuated on end?
		> 20 characters?

Example grouping: Audio extensions

aac abs aif aifc aiff aiffc au aud aup auf caf cda cdda flac m4a m4b m4p mid midi mp2 mp3 mpc mus ogg pcm ram ra snd sndr sndt sng wav wavpcm wma wv playlist soundlist amr awb cmx emelody emy ime imelody imy kws m4r mld mmf mms morse mot motbin nokia noktxt nrt ott qcp rmf rmi rmid rng rtttl rtx sag sagem smaf im_ mlp mvs sfk vag vwp wvc hft ac3 tune s8 s16 s24 s32 u8 u16 u24 u32 w64 ul lu vox amb cvs cvsd cvu fssd lpc lpc10 vorbis sou sox txw wve cs_ ape idd

Example grouping: Security-related directories

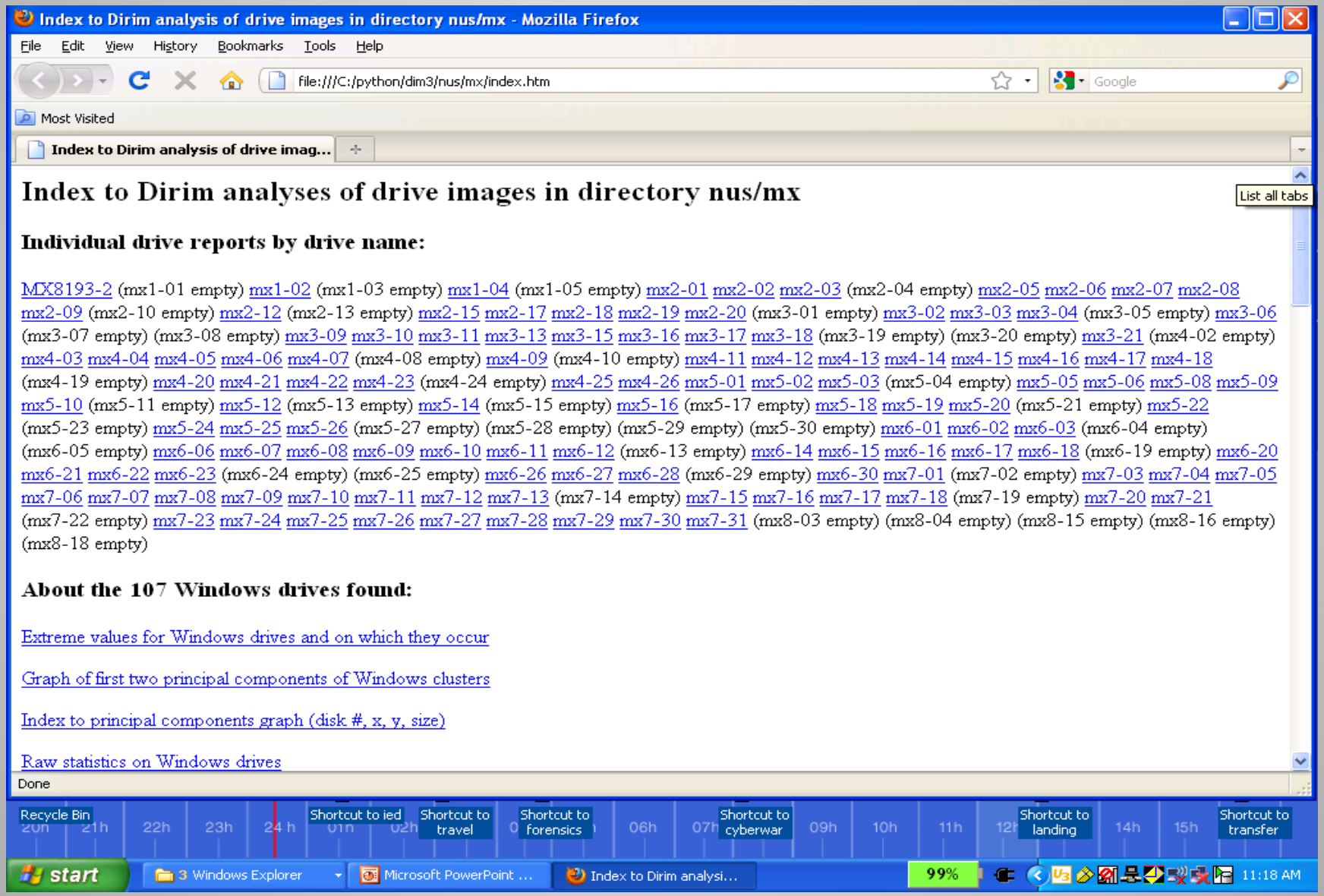
security, sicherung, nprotect, norton internet security professional, norton internet security 2005, norton personal firewall, norton antivirus, norton utilities, norton internet security, nortonav####, norton antivirus corporate edition, norton antivirus #### professional, exitem2080_norton$20internet$20security$20other_1.0_english, norton corporate edition, symantec, symantec shared, symantec antivirus, symantec client firewall, symantec antivirus corporate edition, panda antivirus, panda antivirus 5.0, panda antivirus 6.0, panda antivirus 7.0, panda antivirus 8.0, panda antivirus 9.0, panda antivirus #.#, mcafee virusscan, mcafee for nt(intel), mcafee8.5i, antivirus macafe, keys, certificates, systemcertificates, signatures, firmas, credentials, rav, virusscan, antivirus espaã±ol, antivirus espaÃ±ol, adm, protect, encrypted, compressed and encrypted, antispam, installshield_demoshield_#.##???, installshield_demoshield_#.##esd, quarantine, crypto, rsa, kernel, icsxml, virus defs, ad-aware, spybot - search & destroy, ncdtree, respaldo, respaldo_sga_cc_########, respaldo mrgfs# ######, respaldo 2004 bye indesol, respaldo 2008, grouppolicy, certsrv, norton cleansweep, norton systemworks, installshield, virusscan engine, spyworks v#.#, spyworks63, spyworks##, grisoft, avg#, novell, clr security config, trend micro, savrt, trustlib, intertrust, wuredir, sophos, symcdata, inocit, websrvx, languard network security scanner #, ids-diskless, zonelabs, zonealarm, microsoft antispyware, nailogs, admcgi, pintlgnt, ibm dcm, eacceleration, privacy, virus, virusa, virusi, viruses?, scandisk, scanprog, scanreg, bootscan, vscan, epoagent, virusdef, virusdefs, anti-virus, installshield installation information, esafe, cryptokit, norton, norton 360, pgp corporation, quick heal internet security, net protector, net protector 2005, net protector 2006, net protector 2007, net protector 2008, net protector 2009, net protector 2010, net protector 2011, net protector 2012, checkpoint, quick heal total security, antivir desktop, machinekeys, smartcrypto, keylocker, combofix, steganos.ico, steganos_av.ico, system.security, x86_policy.8.0.microsoft.vc80.crt_1fc8b3b9a1e18e3b _x-ww_77c24773, x86_policy.7.0.microsoft.windows.cplusplusruntime_ 6595b64144ccf1df_x-ww_a317e4b3, spyworks v7.0, spyworks70

File percentages in the Real Drive Corpus

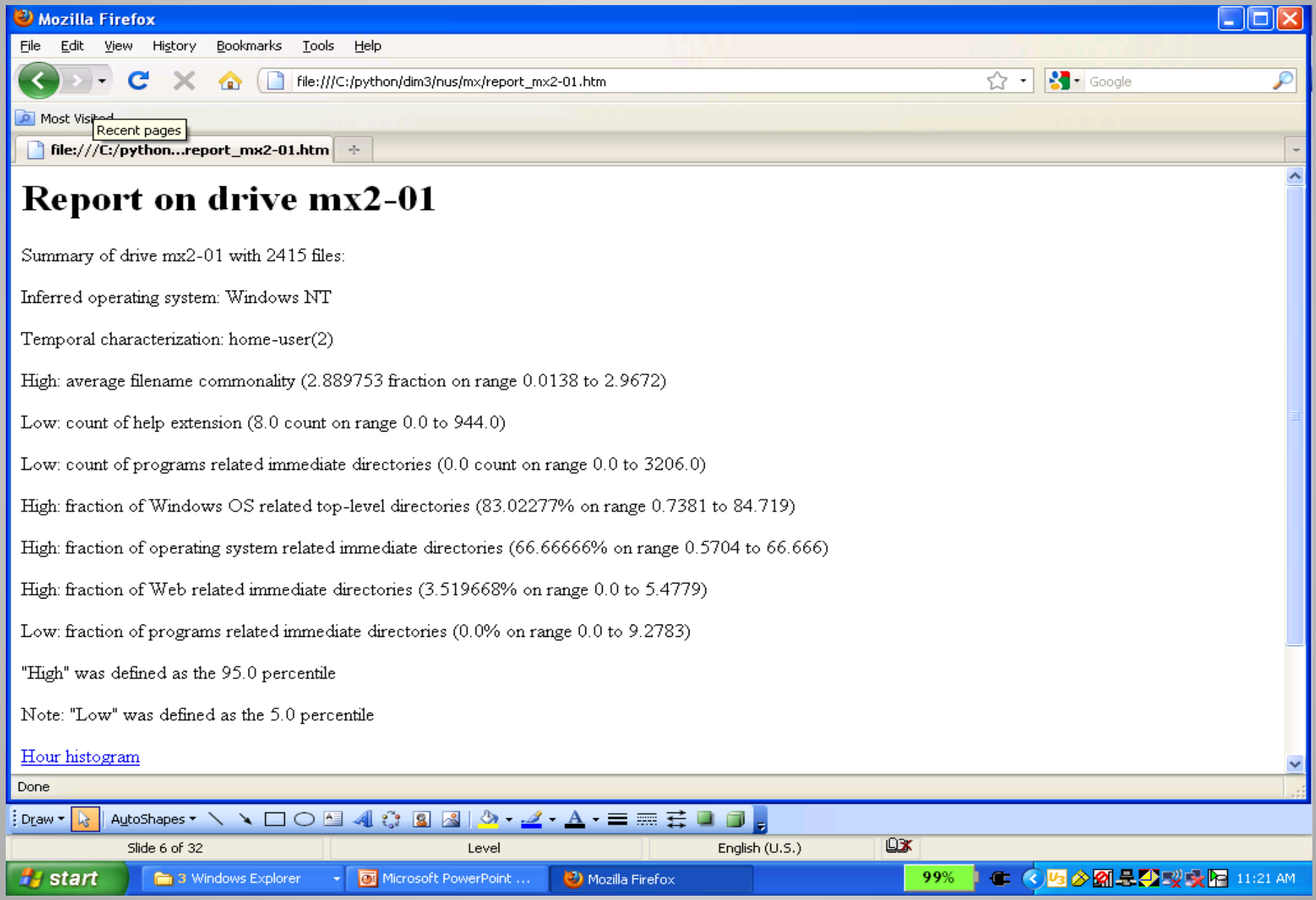
Extension		Graphics	28.8%	None	14.8%	Executable	12.5%
Web	5.6%	Windows	5.3%	Photo	5.1%	Audio	3.6%
Config-urations	3.1%	Game	2.6%	Non-MS document	2.1%	Multiple use	1.7%
Temporary	1.5%	Links	1.2%	Help	1.1%	XML	1.0%
Low frequency	0.9%	Log	0.8%	Program source	0.8%	Microsoft Word	0.7%
Query	0.6%	Spread-sheet	0.6%	Encoded	0.5%	Copy	0.5%
Database	0.4%	Integer	0.3%	Video	0.3%	Security	0.3%
Disk image	0.3%	Present-ation	0.3%	Geogra-phic	0.2%	All other	1.1%
Top directory		Deleted file	27.7%	Program	23.4%	Microsoft OS	19.6%
Document	13.6%	Temporary	4.4%	Unix and Mac	3.8%	Game	3.5%
Hardware	0.7%	Root	0.3%	Microsoft Office	0.1%	Docs. and Settings	0.1%
Immediate directory		Root	25.7%	Temporary	15.3%	Operating system	13.7%
Application	10.1%	Visual	9.8%	Documents	4.6%	Hardware	3.3%
Audio	3.1%	Games	2.0%	Installation	1.5%	Data	1.4%
Help	1.4%	Web	1.3%	Logs	1.2%	Programs	1.1%
Security	1.1%	Sharing	0.9%	Video	0.3%	All other	0.6%

This uses mappings on over 8000 extension and directory names.

Automatically generated report on drives



Report on an individual drive



Fraction of files with NSRL hashcodes in our corpus, by extension group

Group	Fraction	Group	Fraction	Group	Fraction	Group	Fraction
no ext.	.038	Windows OS	.402	Graphics	.449	camera image	.172
temp-orary	.064	Web	.394	General document	.247	Micro. Word	.163
present-ation	.680	database	.193	Other Microsoft Office	.667	spread-sheet	.164
email	.303	link	.063	encoded	.246	help	.724
audio	.125	video	.282	program source	.487	executables	.648
disk image	.346	XML	.482	log	.037	geo-graphic	.136
copies	.228	dictionary	.154	query	.404	integer	.052
index	.030	form	.076	config-uration	.476	update	.015
security	.439	known malware	.016	map	.092	multi-purpose	.372
direct-ory	.503	lexicon	.944	games	.188	engin-eering	.421
science	.505	signals	.088	virtual machine	.135	miscel-laneous	.210

Average for corpus was 0.30. These numbers are disappointing for executables and associated files.

Fraction of files found in NSRL by directory

Group	Fraction	Group	Fraction	Group	Fraction	Group	Fraction
root	.058	operating system	.453	hardware	.579	backup	.150
temps.	.303	help	.669	visual images	.499	audio	.068
video	.185	Web	.412	data	.201	programs	.301
docum-ents	.348	sharing	.310	security	.072	email	.301
games	.115	installation	.345	applications	.382	miscel-laneous	.146

Example software coverage in NSRL

Fraction	Product Name	Count	Fraction	Product Name	Count
0.000	mcafee security scan	252	0.000	the bitmap brothers	4156
0.000	videolan	185455	0.000	xunchitools	1734
0.000	apple software update	2641	0.000	media player classic	137
0.000	informic	1171	0.000	winutilities	549
0.000	hypertechnologies	495	0.001	webacus	871
0..005	infograes interactive	12142	0.091	zone labs	439
0.296	sonysz32audio	517	0.538	macromedia	44570
0.589	openoffice.org.2.2	4798	0.798	microsoft plus!	2329
0.900	norton utilities	178	0.930	roxio	1268

- This gives the fraction of files under that software's directory in our corpus that had hashcodes in NSRL.
- This included software under Program Files, bin, and known top-level software directories.
- Over half the software (out of 4,377) had no hashcodes at all in NSRL. Be aware.

Some software directories with zero coverage in NSRL

o2m30102, o2m30103, o2micro30100, o2micro30101, oam, obama-alien-defense, ocins, ocsoasis, odbc, odesa yazc#4b#1lc#4b#1m, offerapp, office 2007, officeupdate11, ofis2003tr, oggplay, ojosoft, ojvm_g, ojvm, okidata, old song, old_exe, olivetti, omegaone, omrwxqbr, on-line help console, oneroof cybercafepro client, onflow, online tv player 4, online-dienste, onlineprint, online~1, open text, openoffice.org 1.0.2, openoffice.org, opentable, opl3, opl4, opldhuxqn£0™Ëfsvcxlcg.exe, opro, ops647, optimize, orancrypt8.dlli, orange, organizerinstaller, orgplus, orgpub7, originals, orjinal_xp_yapma, oryte_games_1.9, os2, others, ousb2, overdrive, pac-man, pacman2, pacpc, padornew, pager_applet, pagoware, paltalk, pandora.tv, partypoker, pc inspector file recovery, pc print, pcwizard, pdf to epub converter, pdf-convert, pdfcreator toolbar, pdo, pdt, peacemaker, pen drive, pend, performance, personality, persys, petroleum experts, pe, phatpad, photoeditor, photofiltre, photosafe, photos, phpnukeen, pics, picture puzzle.net, pingple2.0, piqclogj, planetzero, player2, player_online, player, playit, plus on, pnevaxgk, poamcpalst&ÖËpwfxuiks.exe, pocket rar, pocket stock monitor, pocket tank deluxe, pocket tanks, pocketdictionary, pocketquran, poc, pointdev, pokemon paint v1.00, pool, popupwithcast, postclie, powerball, powermp3, powerpc, pplive, ppview, precisiontime, preloaders, prity, priyanka, prjclient, process, procman, program files, program shortcuts, projection, prolific, promotomobile, prosetdx, protocol, prtpkt, prtserv, psconvert, psp, pub, puerclient, pumpkin-push, putty, puxwddjt‰5t×Ëkcclxxgp .exe, puzzle-boy, puzzleinlay, puzzle, pvplayer, pylkthwi, pzdialer

Gaps in the coverage of NSRL

Despite 21.0 million distinct hashcodes in NSRL:

- Missed by NSRL: "machine.inf" of size 103496 occurs 40 times in the corpus with the same hash value.
- Missed by NSRL: a 56-byte GIF image under names "BTN-DO~1.GIF", "TB_SRH~1.GIF", "DF_REV~1.GIF", "ICON_A~1.GIF", etc.
- Missed by NSRL: another 56-byte file that occurred 912,013 times in the corpus, usually under the name “.” or “..”.
- Missed by NSRL: cache files like one of size 10 with names “A0021284.ini", "A0021290.ini", "A0022464.ini", "A0022478.ini", etc. under System Volume Information in Windows.

Quick additions to NSRL

- Our corpus allows us to suggest two obvious kinds of additions to NSRL hashcodes:
  - Hashcodes that occurred on more than 5 drives;
  - Hashcodes on files with the same pathname, minus the extension, as files with NSRL hashcodes.
- Altogether we found 937,570 additional hashcodes using these rules on the 45 million files of the corpus.
- We are currently researching other sources of hashcodes like hashsets.com.

Counts of overlap between hash sets

For hashcode in row, count in column	Corpus	NSRL	Occurred at least 5 times in corpus	Same path as corpus hashcode in NSRL	Hashsets .com
Corpus	9,098,822	465,209	207,209	729,411	301,407
NSRL	465,209	21,043,342	7	8	879,769
Occurred at least 5 times in corpus	207,209	7	208,789	58,534	3,176
Same path as corpus hashcode in NSRL	729,411	8	58,534	1,179,203	19,013
Hashsets. com	301,407 (260,199 unique)	879,769	3,176	19,013	6,441,457

Example filename discrepancies: Corpus vs. NSRL

Corpus	NSRL	Corpus	NSRL
..	scriptsIcon.png	afd.sys	afd.sy_
mflm.in_	mflm.inf	iewebhlp.chm	iewebhlp.ch!
hostconfig~	hostconfig	#139#bldtips.cst	_139_bldtips.cst
palm tree.bmp	palmtree.bmp	gnome-text-x-c++.png	gnome-mime-text-x-c++.png
fireworks.pot	FIREWORK.POT	customer.dbf	customers.dbf
adobebannereng.awe	AdobeBannerenu.awe	lsen40es.hlp	LSEN40EN.HLP
trofeo.wmf	trophy.wmf	default_ns_2.css	default_nss.css
market.ini	IMPORTANT.GIF	graphic.xfo	graph.xfo.70DBED24_B579_...
0000007c.query	__0X023F	displaylanguagenames.gv_gb.txt	DisplayLanguageNames.gv_GB.t
netscape.cfg	netscape.cfg.htm	wmiadap.exe.new	wmiadap.exe
_er7b2~2.tmp	ajbs	nonet.html	nonet.html55
audit.chm	auditw.chm	nerodigitalext.dll	NeroDigitalExt1737449D.dll
desktop_icon_01.bmp	HCimgE40.bmp	memo wizard.wiz	MEMO.WIZ_1033

Discrepancies in file names: NSRL vs. corpus

Type of inconsistency	Count on corpus
Missing extension in NSRL	103,036
Missing extension in corpus	18,905
Additional 128-bit hash code on NSRL extension	19,959
Additional backup extension on corpus item	18,330
Other additional extension on NSRL item	2,814
Other additional extension on corpus item	2,311
More detailed extension on NSRL item	273,558
More detailed extension on corpus item	9,835
Same file name with complex difference in extension	19,900

Discrepancies in filenames: NSRL vs. corpus (2)

Type of inconsistency	Count on corpus
NSRL file name has numeric addition	2,932
Corpus file name has numeric addition	4,983
Corpus file name same except has a bracketed number	16,536
Only difference in file names is punctuation	6,102
Apparent misspelling in NSRL file name	5,550
Period on end of NSRL file name	280
"__OX" code in lieu of a NSRL file name	169,708
Corpus file name is a placeholder like “.”	1,056
Plural in NSRL and singular in corpus	139
Plural in corpus and singular in NSRL	359
Foreign-language version identical in contents to the English version	67

Discrepancies in filenames: NSRL vs. corpus (3)

Type of inconsistency	Count on corpus
Exclamation subsitution at end of NSRL file name	42,267
Exclamation substitution at end of corpus file name	0
Underscore substitution at end of NSRL file name	351,989
Underscore substitution at end of corpus file name	4,938
NSRL file name is abbreviation of corpus	35,698
Corpus file name is abbreviation of NSRL	1,600
Same extension but NSRL file name more detailed	117,292
Same extension but corpus file name more detailed	17,950
Match of files under 100 bytes	71,917
Files occurring with more than three names in corpus	19,706
Significant differences in file name and extension	277,231

No evidence for errors in NSRL hash codes

- One approach: Find hash codes occurring at least 5 times in the corpus whose hash code never occurred in NSRL but whose name occurred in NSRL. Result: none found.
- Another approach: Look for hash codes occurring at least 5 times in the corpus but never under the NSRL-given filename (43,384 total).
  - 19,028 cases:The NSRL name was the same except for a final underscore or exclamation point
  - 903 cases: The corpus name was embedded in the NSRL name
  - 1,535 cases: The NSRL name had additional characters
  - 111 cases: Additional minor differences
  - 4,363 cases of the same extension with very different file names
  - 15,500 remaining cases (better than 277,231): Both the extension and filename did not match. 80% were deleted files, normally 32% of the corpus, which suggests many were Sleuthkit errors on deletions.

Discrepancies in file sizes: NSRL vs. corpus

- NSRL-reported file sizes differed on 7,461 corpus files, and usually by large amounts.
- Example: test2.zeros of size 4096 in NSRL which had the same hash value as file AA00389B.71 of size 2490544 in the corpus.
- Possible explanations:
  - This is a hash collision, unlikely for the size of the SHA-1 hash space.
  - The hash values could be incorrect. This is unlikely when (usually) names matched.
  - NSRL may be measuring a block size since errors are all powers of 2. This is unlikely since NSRL sizes appear otherwise reliable.
  - The file sizes that SleuthKit retrieved for some unallocated corpus files were incorrect. This explanation appears to be the most likely.

Product codes in NSRL

- We classify files in our corpus, but the philosophy of the NSRL "product codes" is different: They describe the package in which the file came, not the file itself.
- So Program Files/Microsoft Office/media/CntCD1/Photo1/j0180794.jpg is classified as an image file by our taxonomy (by both extension and directory) but as applications software by NSRL.
- Beware the "hacker tool" category in NSRL since it is very incomplete.

Conclusions

- The NSRL RDS has good precision, but its recall is imperfect and users should be aware of that.
- Coverage was over a wide range of categories, and was not just confined to executables.
- 74% of software in our corpus were substantially uncovered by NSRL, so it has definite gaps.
- We found errors and possible improvements for NSRL file names.
- Comparison of our corpus with NSRL did pinpoint some errors in our own drive imaging on file sizes.
- Simple additions to the NSRL RDS could improve its coverage by a million hash values.
- NIST's approach of not running software is unreasonable with today's complex software installations.
- Our Dirim suite of tools and analysis results are freely available for research.