AV Testing Exposed

Peter Košinár, Juraj Malcho, Richard Marko, David Harley

{kosinar, malcho, marko}@eset.sk, dharley@eset.com

Testing? What the heck?

Joe: Hi, I'd like to buy a good antivirus.

Shopkeeper: Great! Here, they're all on that shelf.

Joe: Oh my, it's full of stars! Which one shall I take?

Helpful tester A: I'd recommend X, it's the best!

Helpful tester B: Take Y, it's number 1 in my test!

Helpful tester C: Nah, trust me, it's Z!

Other helpful testers: Me, me, me!

Joe: Darn, so many choices and so little time…

Why testing?

Pro: Testers want to help

- inexperienced public (decisions),
- AV companies (feedback),
- themselves (money, fame, …).

Con: Tests are often inconsistent, inaccurate, not repeatable, not verifiable, … utterly useless(?).

Testing problems

Bad methods = bad results, good methods != good results.

AMTSO guidelines are **baseline**, not **ultimate goal**.

Conflating testing and certification.

Mixing apples and oranges into a low quality fruit salad.

Performance testing? Hard to do, easy to fool.

Detection testing? Even worse!

How we see detection testing

1) Assemble test-set
2) Run products
3) Grade them
4) Rinse and repeat

The test-set

Contents of the test-set decide the result, GIGO.

Data sources?

Representing the intended set? Not more? Not less?

Size matters, but validation even more so!

Repeatability and independent verifiability?

The test-set – Wildlist based?

Good: Very well tested, valid samples.

Bad: Very small, not representative of the real threats.

Good: Repeatable, available for independent review.

Bad(?): Suitable for Pass/Fail, not 47.2% tests.

Good: Replication is hard to fool.

Necessary conditions

Source Neutrality

One source – bias, many sources – difficult merge

Reputation/prevalence helps, but is not silver bullet.

Validation

Do the samples actually work? How do you know?

Do they belong into this test?

Diversity

Are you sure nobody is over- or under-represented?

How do you know without doing full RE?

Size does matter!

Too few samples

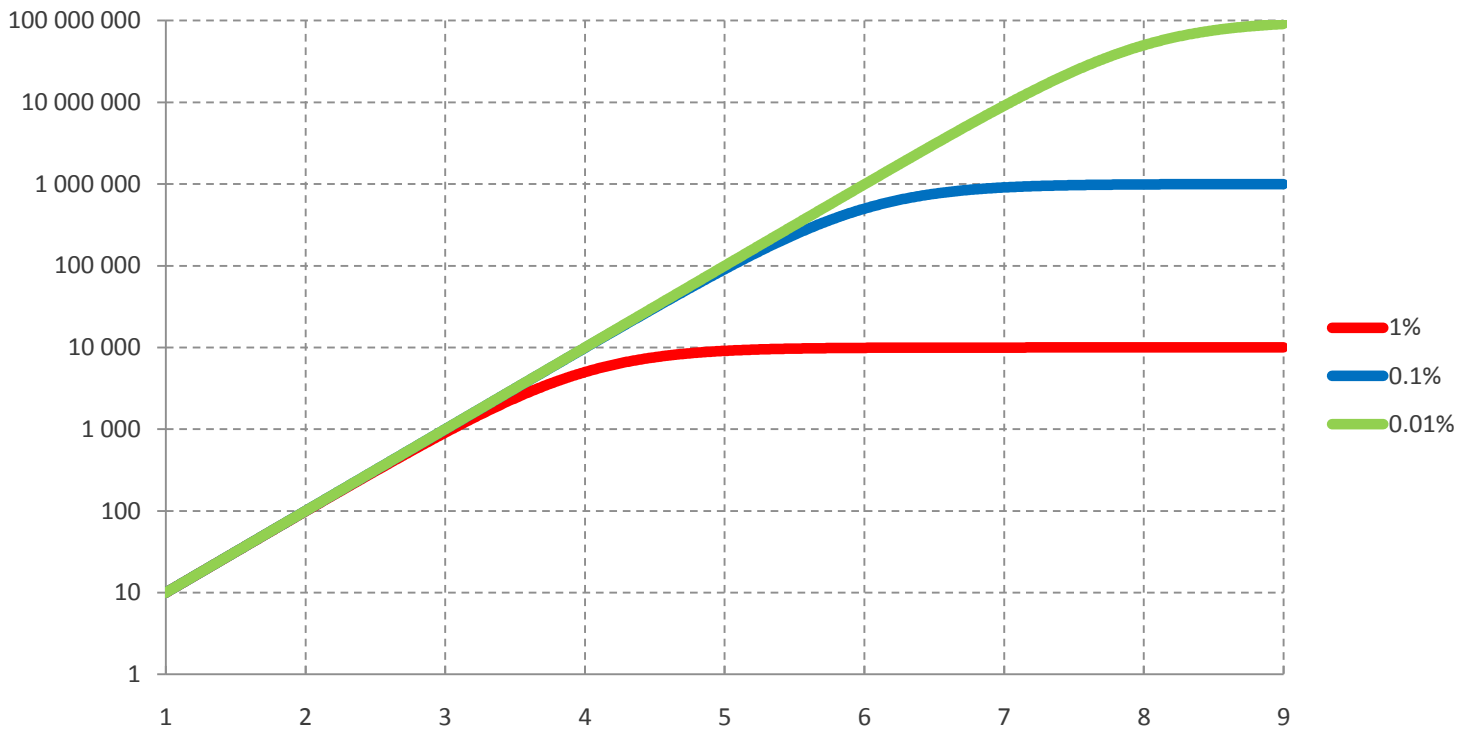
Small fraction of samples used = large error. How large? Standard deviation for 1 million full set and 80% actual detection rate:

10	20	50	100	1000
12.7%	8.9%	5.7%	4.0%	1.3%

And for 60% detection, it's even more volatile:

10	20	50	100	1000
15.5%	11.0%	6.9%	4.9%	1.5%

Too many samples



Rule of thumb: Each digit requires 100-times larger set.

Validation and diversity

Multi-scan promotes engine-sharing

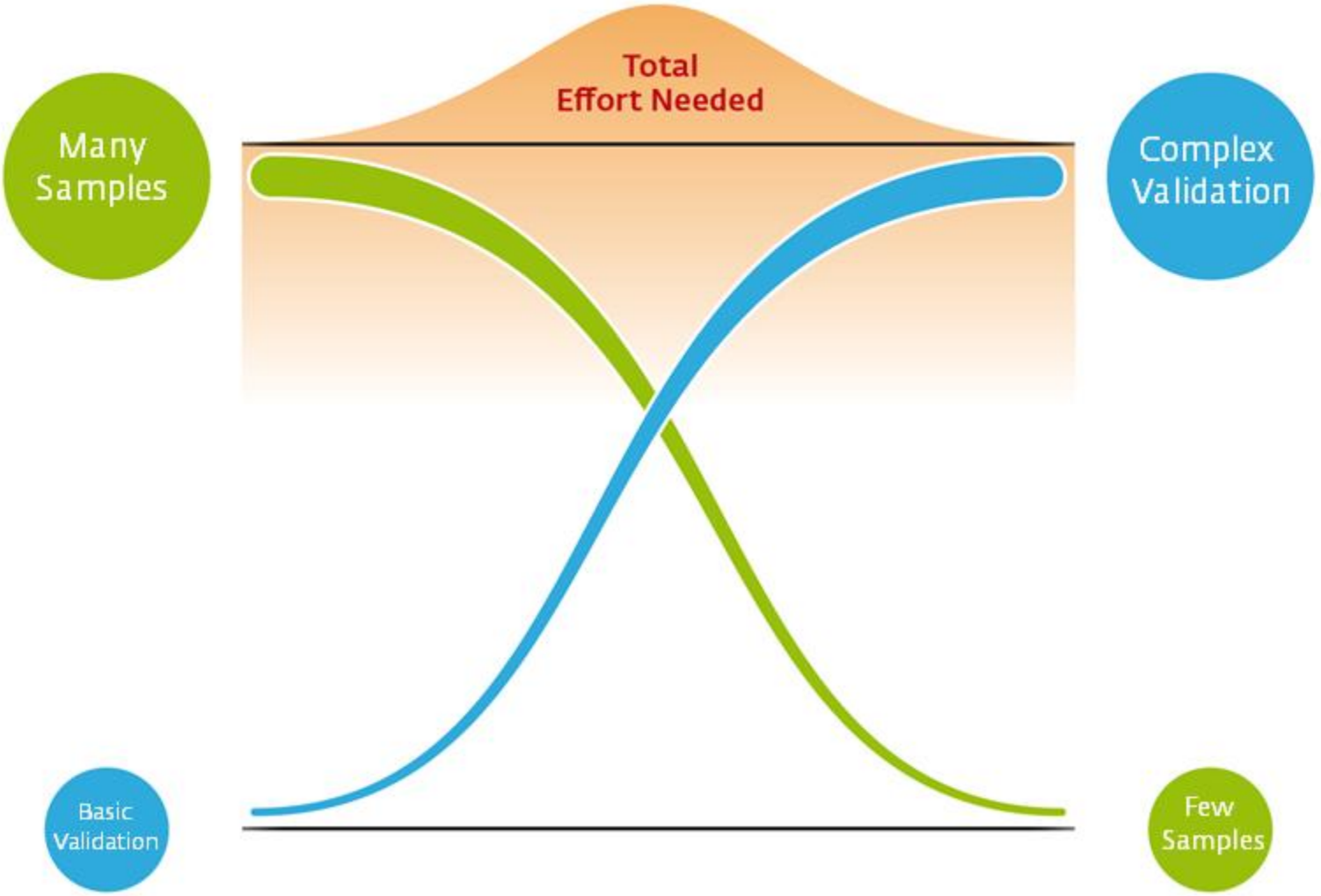
One engine gets more “votes” than another.

Inappropriate files reward junk-detection

Detecting files you don’t care about can help you in test…
... but not anywhere else. 8% can turn the tables badly:

Product	Test result	“Junk”	Reality
Innocent	91.88%	10%	99%
Guilty	93.80%	80%	95%

Number of samples vs. validation quality



AV companies under a DoS

Testers lack the capacity to check their data

Hey, AV company, check this for us…
After all, it's only fair that you'll spend ~150 man-days every couple of months just to get the result you deserve.

Customer is the king

Magazines and people asking for artificial tests are granted their wishes, however foolish they are.

Fruit salad

Jamie Olivier would be envious to see such a mix of apples and oranges.

Cooperative or death-match?

Innovative approach

“I know something I won’t tell”. You want to take a look?
Pay! See also: Pig-in-a-poke.

Nah, you can’t see the data, we don’t trust you.

So you don’t, but we should?

Conclusion

Samples selection is crucial

Failing to reflect the accuracy of samples selection in the results renders the whole test irrelevant.
Transparency is a must – otherwise the test has no meaning.

AMTSO doesn’t have answers to everything

Its documents help, but they’re not a crystal ball.

Faulty tests are putting AV companies under DoS

And inflict damage instead of helping anything.
Bonus: Misinformed and mis-educated users.

Questions?

Peter Košinár, kosinar@eset.sk
Juraj Malcho, malcho@eset.sk
David Harley, dharley@eset.com
Richard Marko, marko@eset.sk