KASPERSKY lab

Machine learning et expertise humaine

« Comment les algorithmes avancés de Kaspersky Lab garantissent-ils la meilleure protection pour votre entreprise contre les cybermenaces ? »

www.kaspersky.fr

#truecybersecurity

Machine learning et expertise humaine

Chaque année, la qualité de notre protection Kaspersky Lab remporte plus de prix que quiconque dans le secteur de la cybersécurité. Cet exploit serait impossible sans notre veille HuMachine™ propriétaire : la fusion d'un cybercerveau de « big data » mondial s'appuyant sur les algorithmes du machine learning et de l'expertise inégalée de nos équipes de sécurité dans la lutte contre les menaces de nouvelle génération.

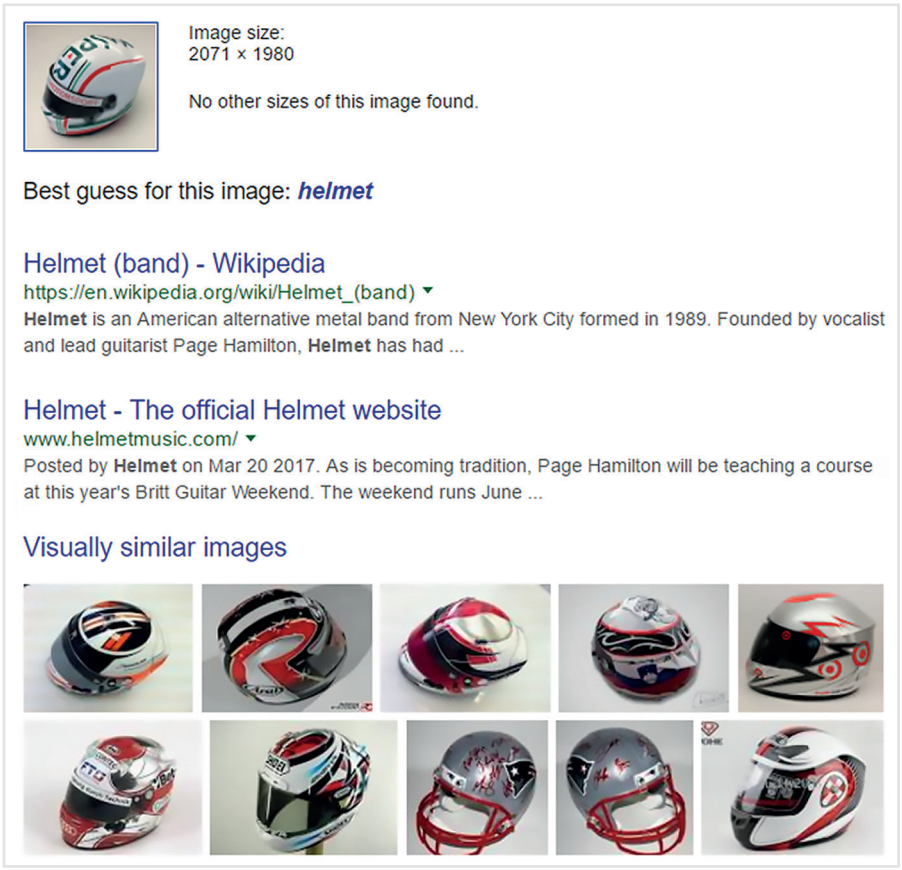
Nous vous offrons une « immersion » au cœur de l'infrastructure anti-programmes malveillants de Kaspersky Lab, révélant nos algorithmes et leur rôle dans la lutte contre les menaces les plus dangereuses pour les entreprises comme la vôtre.

« Comment les algorithmes avancés de Kaspersky Lab garantissent-ils la meilleure protection pour votre entreprise contre les cybermenaces ? »

Notre approche traditionnelle de détection automatique

Notre base virale contient des échantillons de virus détectables regroupés par noms de détection, par exemple Backdoor.Win32.Hupigon.abc. Quand un nouvel échantillon non détecté arrive, nous commençons par rechercher des échantillons similaires dans notre base virale. Le principe de recherche est à peu près le même que celui utilisé par Google Search. La seule différence repose sur le fait que Google Search s'appuie sur un mot, alors que nos recherches sont basées sur les caractéristiques de fichier. Dans le scénario le plus simple, si l'échantillon est décompressé correctement, il est possible d'extraire les chaînes responsables de la fonctionnalité du programme malveillant et de les utiliser presque de la même façon qu'un moteur de recherche utilise les mots clés.

Au sein de Kaspersky Lab, nous avons un système automatisé qui gère à la fois l'analyse des fichiers et le classement automatique des menaces.



Recherche d'images similaires sur Internet par le biais du service Google

Ce système trie les flux entrants d'échantillons tout en ajoutant des hachages pour identifier et définir les détections. Un dossier simple de hachage couvre la détection d'un seul fichier, mais de cette façon, nous pouvons être sûrs qu'il n'y aura pas de « faux positifs ».

Lorsque la base virale ne contient aucun échantillon similaire au programme malveillant, nous savons qu'il s'agit de quelque chose de totalement nouveau, ou qu'il ne s'agit pas d'un programme malveillant. C'est là que l'expertise des analystes spécialisés en antivirus entre en jeu. À la décompression et à la détection d'un échantillon, l'analyste crée une sorte de « centre de gravité » dans la base virale. Au fil du temps, des versions modifiées du nouvel échantillon vont automatiquement graviter vers ce point de référence.

Approche heuristique de détection automatique

Seule la détection basée exclusivement sur les hachages permet d'arriver jusque-là : une légère modification du fichier (un seul octet ajouté à la fin, par exemple), et le fichier entier devient à nouveau indétectable. C'est pourquoi nous lançons notre système de détection automatique heuristique sur l'ensemble de la famille de nos échantillons de programmes malveillants, par exemple, Backdoor.Win32.Hupigon.abc. Avec l'aide d'un émulateur, le système heuristique crée des journaux d'exécution de tous les échantillons, trouve leurs modèles d'exécution courants et crée un seul dossier heuristique basé sur l'exécution. Grâce à cette approche, les nouveaux échantillons de programmes malveillants qui présentent un comportement similaire seront détectés, même si le contenu est quelque peu modifié.

Examinons de plus près le procédé qui permet de créer des dossiers de détection heuristique. Le système robotique utilise le machine learning pour extraire les séquences d'exécution clés. La machine ne connaît pas ou ne s'intéresse pas à l'usage particulier des séquences de commandes. Il lui suffit de savoir que telle ou telle séquence d'exécution, ou combinaison de séquences, est caractéristique de certaines familles de programmes malveillants et qu'elle ne peut pas se produire dans les fichiers sains. Après quelques versions, les indicateurs les plus efficaces, ainsi que leurs combinaisons, sont automatiquement regroupés dans des dossiers.

Contrairement à ce robot, un analyste humain expérimenté comprend exactement ce que l'échantillon prévoit de faire, malgré ses efforts visant à déjouer l'émulateur du système heuristique. Il peut ainsi créer immédiatement un dossier en mettant en évidence le comportement apparent du programme malveillant.

Ces deux approches ont tendance à fonctionner de pair, en particulier lorsque les résultats de la détection automatique ne sont pas probants et qu'un deuxième avis d'expert est nécessaire. Les dossiers créés par l'humain et la robotique fonctionnent alors en tandem, garantissant une meilleure détection grâce à une harmonie HuMachine™ parfaite.

```
KERNEL32!LoadLibrary(0x004020B6 "KERNEL32.dll");
KERNEL32!GetTickCount();
KERNEL32!LoadLibrary(0x00403000 "kernel32.dll");
KERNEL32!LoadLibrary(0x0040302C "urlmon.dll");
urlmon!URLDownloadTofile(,0x00403061 "http                                    ",0x004030C5 "c
KERNEL32!Sleep()
KERNEL32!DeleteFile(0x004030C5 "c:\\boot.bak");
urlmon!URLDownloadTofile(,0x0040308F "http                                    ",0x004030B9 "c:\\4
```

Journal d'exécution Trojan-Downloader.Win32.Small.aon

Pour éviter d'être détecté, le malfaiteur peut modifier les fonctionnalités de son programme malveillant. Mais il existe des limites. Supposons que le programme malveillant dispose de fonctionnalités de base : Télécharger un fichier via un lien malveillant, Enregistrer le fichier sur le disque dur et le démarrer (téléchargeur de chevaux de Troie). Il n'existe pas plus de 10 méthodes de programmation différentes pour télécharger un élément à partir d'Internet et pas plus de cinq pour démarrer un fichier exécutable. Lorsque le malfaiteur les a toutes essayées et qu'il a constaté que chaque méthode est détectée, il est probable qu'il abandonne et qu'il s'attaque à une entreprise ne disposant d'aucune solution de sécurité, ou disposant d'une solution sans outils d'analyse d'exécution.

Le malfaiteur peut avoir plusieurs tours dans son sac. S'il connaît les spécificités de l'émulation, il peut essayer de perturber le processus d'émulation en insérant par exemple un long délai d'exécution ou en demandant des paramètres système que l'émulateur ne peut fournir. Certaines de ces astuces peuvent elles-mêmes être traitées comme des indicateurs de détection, mais nous pouvons néanmoins détecter la véritable fonctionnalité de l'échantillon par le biais d'une autre méthode, à l'aide de la surveillance du système, un système qui surveille les activités d'un processus avec le système d'exploitation actuel.

Surveillance du système et détection de comportement

Contrairement à l'émulateur, la surveillance du système est un véritable système de détection de comportement basé sur les journaux d'exécutions d'échantillons réels, de sorte qu'il est impossible de tricher. Elle possède son propre ensemble de dossiers de comportement qui, à bien des égards, ressemble à celui du système de détection qui s'appuie sur l'émulateur.

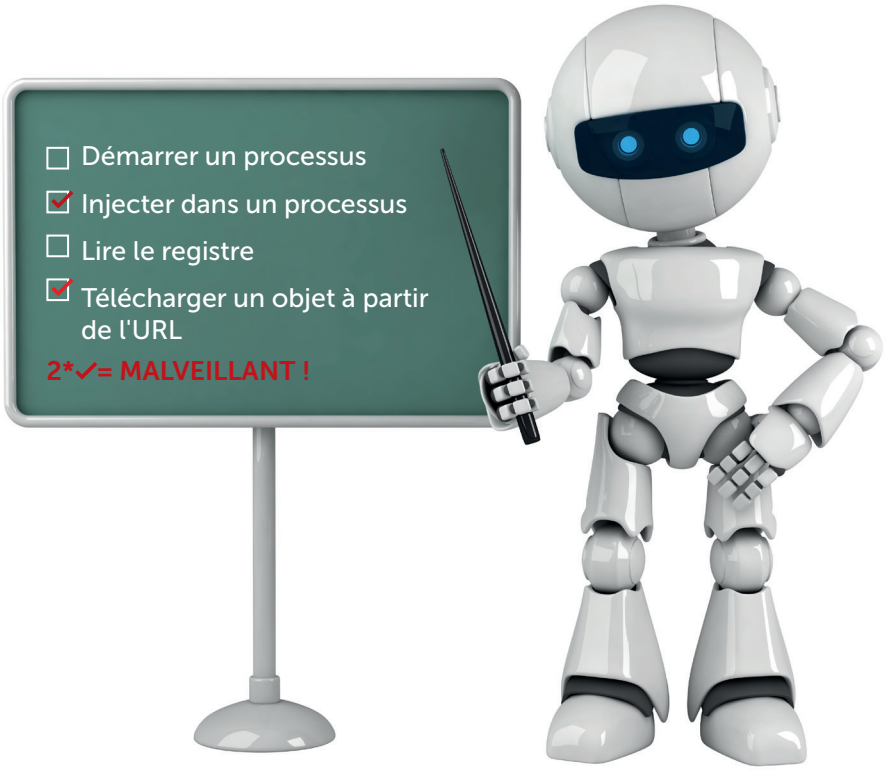
La consignation effectuée lors de la surveillance du système va nettement plus loin que ce qui est possible pendant l'émulation. Et, contrairement au dernier processus mentionné, cette consignation a des délais illimités : tout ce qui semble suspect dans un contexte donné est pris en compte et mis en cache jusqu'à ce que suffisamment de preuves ait été recueillies pour la détection. Si une activité malveillante est détectée, l'action est tout simplement annulée.

Comme avec le système d'émulation, la surveillance du système a un rôle à jouer dans la détection sur site et dans la magie de notre laboratoire. Par ailleurs, l'activité de la surveillance du système est transparente et n'a aucune répercussion négative sur le processus surveillé.

Des analyses de comportement sur site continues créent une couche de détection extrêmement puissante, mais en libérant la puissance de l'infrastructure de Kaspersky Lab pour exécuter les fichiers suspects, étudier leur comportement et fournir une détection de menaces via le réseau KSN (Kaspersky Security Network), nous obtenons un outil encore plus efficace.

Sandbox, KSN et... des hommes.

Comme le suggère notre approche HuMachine™, nous testons continuellement des échantillons, reconnus malveillants et non reconnus, dans nos systèmes Sandbox de comportement interne. Certaines Sandbox imitent les systèmes utilisateur exécutant des produits standard, tandis que les plus puissantes sont dotées de capacités de consignation extrêmement granulaire, permettant une détection extrêmement affinée.



Ajout de comportements suspects révélateurs de comportements malveillants

La consignation Sandbox, ainsi que les statistiques d'exécution de la surveillance du système reçues des participants volontaires du KSN, sont traitées à la fois par les robots et par des experts. Les robots traitent deux processus importants : les journaux de l'exécution de nouveaux échantillons malveillants sont étudiés à l'aide du machine learning pour trouver de nouveaux indicateurs de détection ; des échantillons également inconnus sont détectés, avec les dossiers statiques créés pour une utilisation ultérieure à la fois en laboratoire et dans les locaux des clients. Ainsi, même si les créateurs de programmes malveillants sont assez doués pour éviter la majorité des couches de détection sur site, généralement grâce à des tests de reconnaissance et des tests préliminaires poussés, ils ne seront pas davantage arrêtés dans leur action.

Pendant ce temps, à l'aide d'indicateurs distillés par robot, les experts créent des dossiers de comportement similaires à ceux fondés sur l'exécution émulée, mais avec un éventail beaucoup plus large d'indicateurs à utiliser.

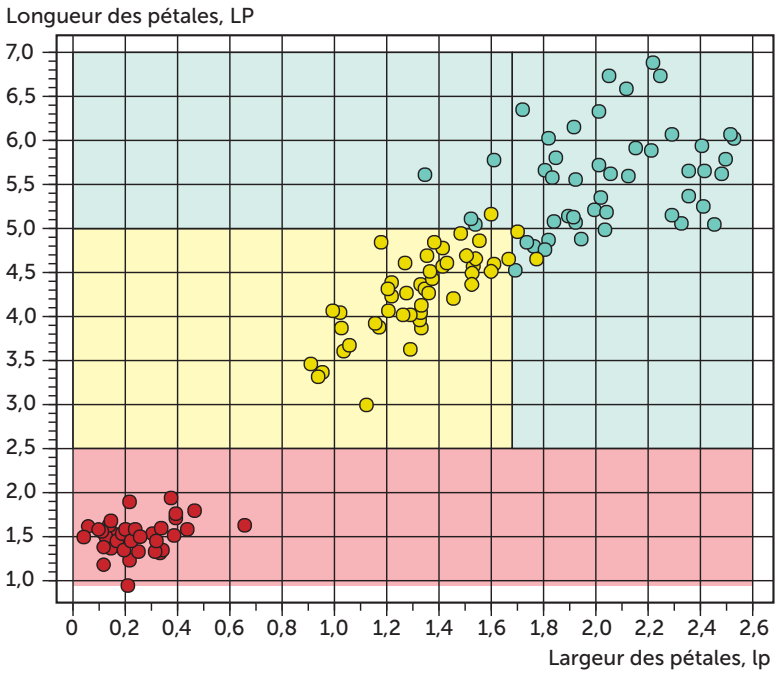
Dossiers intelligents

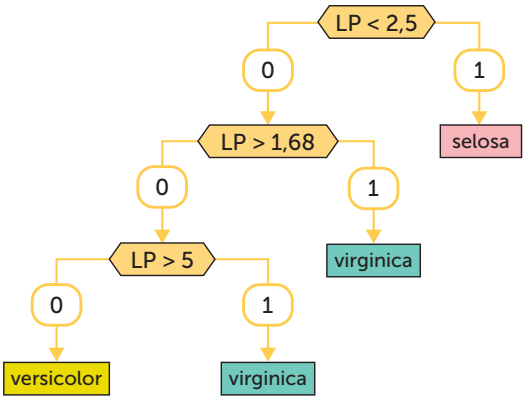
La liste des processus fondés sur le machine learning ne prend pas fin avec ce qui précède. Il existe d'autres couches de détection robotique capables de détecter des familles de programmes malveillants importantes. En règle générale, nous les appelons les « dossiers intelligents ».

Dossiers antivirus basés sur les arbres de décision

La partie robotique du laboratoire de ce système analyse la même base d'échantillons que celle ci-dessus, et crée ou améliore les dossiers en fonction des arbres de décision. Cette analyse permet alors de séparer les fichiers en classes et de spécifier des critères sensibles aux caractéristiques de ces fichiers.

Comment ça marche ? Voyons un exemple basé sur un ensemble de données sur la fleur Iris, un cas de test type pour les techniques de classification statistique. Disons que nous disposons de 150 fleurs : 50 échantillons d'Iris setosa, d'Iris virginica et d'Iris versicolor. Pour simplifier la tâche, prenons les deux caractéristiques les plus instructives de ces fleurs : longueur de pétale (LP) et largeur de pétale (lp). L'identification des caractéristiques de chaque échantillon fournit des données qui peuvent être utilisées pour créer un arbre de décision, qui permet ensuite d'attribuer une des trois classes à chaque nouvelle variété d'Iris par « Requête/Réponse », comme ceci :





Dans le tableau : Axes des 2 attributs les plus instructifs (sur 4) deux classes divisées avec précision, 3 erreurs en classe 3.
Source : Coursera/Yandex

Notre moteur antivirus utilise exactement le même genre d'arbre. Chaque arbre de décision est soigneusement ajusté et remis à l'utilisateur. Les caractéristiques sélectionnées d'un fichier individuel exécuté sur l'ordinateur de l'utilisateur sont extraites et acheminées à travers chaque arbre de décision. L'arbre utilise ensuite les réponses pour déterminer si le fichier est malveillant ou non.

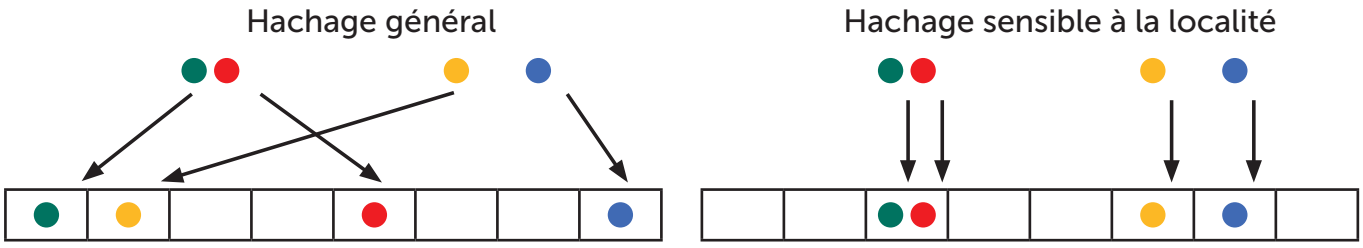
Cette approche a pour avantage de généraliser les capacités. Chaque arbre est créé en laboratoire à partir d'un petit sous-ensemble des échantillons dont nous disposons, mais sur les ordinateurs des utilisateurs, l'arbre détectera également tout échantillon qui n'a pas encore été acquis par notre laboratoire. Par exemple, dans la photo ci-dessus, chaque point de la zone rouge sera détecté comme étant un Iris setosa. Un seul dossier basé sur un arbre remplace en moyenne un millier de dossiers de hachage.

Le machine learning est indispensable à la création d'arbres de décision. Bien qu'ils puissent fournir de longues listes de caractéristiques au robot, les experts ne créent pas personnellement de dossiers basés sur les arbres. Seule une machine peut extraire et appliquer les données, en sélectionnant les meilleures caractéristiques et surtout, en créant des règles de décision basées sur ces caractéristiques. L'expert surveille simplement le résultat et contrôle le processus.

Hachage sensible à la localité

Les modèles de détection à base d'arbres sont très utiles mais ils présentent encore un défaut important. Même s'ils sont créés automatiquement en laboratoire, ils ne fonctionnent correctement que sur l'hôte (l'ordinateur de l'utilisateur) où le fichier en question est étudié. Un système de Cloud basé sur ce principe créerait un trafic réseau considérable, ce qui n'est pas souhaitable dans la plupart des cas.

Les systèmes de Cloud basés sur le hachage, par contre, sont beaucoup plus légers en termes de trafic. Mais un hachage cryptographique classique, tel que MD5 ou SHA256, correspond presque toujours à un seul fichier. C'est une bonne chose de ne pas trouver un second fichier avec le même hachage ; les faux positifs ne sont pas étudiés ici. Mais un hachage identique pour tous les programmes malveillants d'une même famille serait utile. En d'autres termes, les modifications de fichiers insignifiantes n'auraient pas d'incidence sur le hachage. En réalité, le LSH (Locality-Sensitive Hashing), ou Hachage sensible à la localité, le permet. Les requêtes qui engendrent des détections basées sur ce hachage peuvent être exécutées via le Cloud.



Les points multicolores (fichiers) sur la gauche sont hachés à l'aide de l'approche traditionnelle. Les hachages n'ont rien en commun. À droite, ils sont hachés à l'aide du hachage LSH. Les fichiers qui ne sont pas très différents obtiennent des hachages identiques.
Source : 0110.be

Comment calculer les niveaux de similitude entre les fichiers ? Voici un exemple intéressant :

Supposons que le fichier A est défini par les caractéristiques numériques suivantes :

31, 83, 98, 86, 183, 79, 67, 153, 77, 67

Prenons ensuite le fichier B qui est légèrement différent :

27, 89, 93, 81, 190, 71, 67, 161, 75, 69

Tous les numéros peuvent être « arrondis » au nombre inférieur en les divisant par 10. Nous obtenons alors :

Fichier A : 3, 8, 9, 8, 18, 7, 6, 15, 7, 6
Fichier B : 2, 8, 9, 8, 19, 7, 6, 16, 7, 6

Comme vous pouvez le constater, les valeurs des caractéristiques sont maintenant presque identiques.

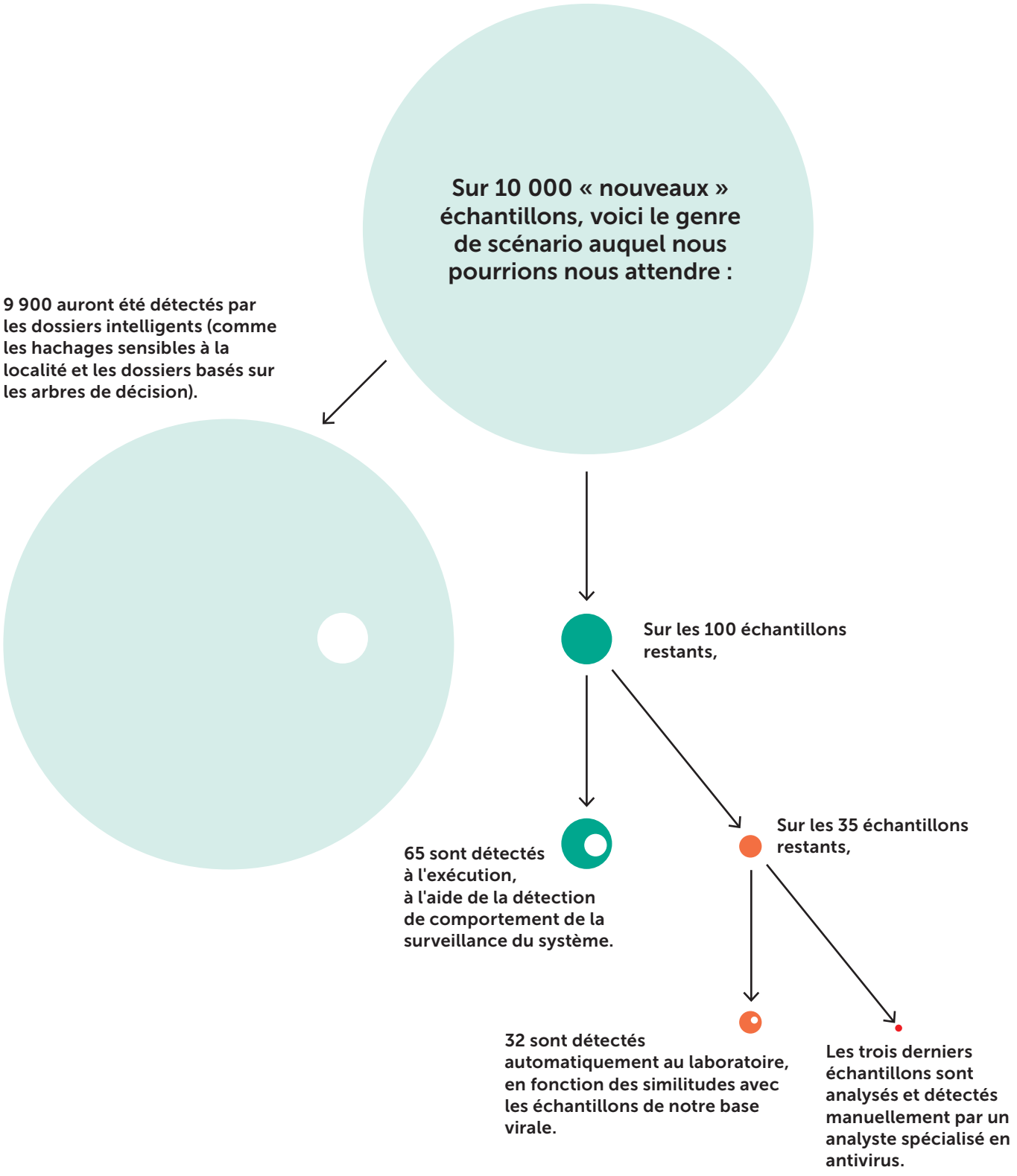
Voici une autre approche : calculer la moyenne arithmétique des nombres dans la première et la deuxième moitié de chacun des deux fichiers ci-dessus. La réponse devient alors :

Fichier A : 96, 88
Fichier B : 96, 88

Dans ce cas, les hachages LSH sont identiques.

Le défi de cette approche implique de choisir des caractéristiques qui varient légèrement à l'intérieur d'une même famille de programmes malveillants, mais qui sont encore suffisamment différentes pour être reconnues lorsqu'un fichier sain est concerné. Ces caractéristiques doivent alors être « quantifiées » (autrement dit, elles sont traitées de façon à réduire leur précision). Comme vous l'avez peut-être deviné, seul un robot peut le faire. Mais la tâche est encore élaborée par un expert.

Chemin du programme malveillant



Tous les échantillons (quelle que soit la façon dont ils le sont dans la base virale) sont fréquemment analysés de nouveau pour chaque nouvelle détection à l'aide de technologies de généralisation (dossiers heuristiques automatiques/dossiers basés sur les arbres/dossiers sensibles à la localité décrits précédemment).
Si un échantillon a déjà été détecté uniquement à l'aide du hachage individuel, la détection est « généralisée » par le machine learning, de sorte de l'inclure dans une grande « famille » de programmes malveillants décrite uniquement par un seul dossier. Le dossier de hachage individuel est ensuite supprimé.

Écarter les faux positifs

L'histoire de la détection heuristique par le machine learning ne serait pas complète sans évoquer la question des faux positifs. Comme avec toutes les méthodes basées sur le principe de généralisation, ces techniques contiennent une probabilité inhérente d'erreur, qui donne lieu à la détection de faux positifs. Les changements imprévus dans le paysage de la menace peuvent augmenter cette probabilité. Par conséquent, un réglage constant des modèles de détection et un contrôle très poussé et constant des faux positifs sont nécessaires.

Les produits Kaspersky Lab intègrent des mécanismes automatisés pour le suivi, la désactivation rapide et la correction des dossiers défectueux. Cependant, partant du principe que la multicouche est présente partout et pour garantir le meilleur résultat possible aux clients, tous les dossiers, y compris ceux créés par des robots, sont constamment surveillés par les analystes les plus expérimentés. Ils s'assurent que les dossiers sont soigneusement testés et ajustés régulièrement pour garantir les plus hauts taux de détection tout en conservant un nombre de faux positifs le plus près possible de zéro. Comme les essais indépendants le prouvent sans cesse, ils excellent en la matière !

Toutes les technologies et approches décrites ici ont un rôle décisif dans l'efficacité des mesures de cybersécurité, mais nous sommes toujours en quête de nouvelles méthodes, pour maintenir à distance chaque nouvelle génération de menaces.

Tout savoir sur la sécurité sur Internet : **www.viruslist.fr**
Rechercher un partenaire près de chez vous :
http://www.kaspersky.fr/partners/buyoffline/liste-des-partenaires

www.kaspersky.fr