大数据情境行为分析要求

非官方中文译文·安天技术公益翻译组 译注

文档信息			
原文名称	THE REQUIREMENTS FOR A CONTEXTUAL BIG DATA BEHAVIORAL ANALYTICS ENGINE		
原文作者	STEVE KING	**原文发布日期**	2016 年 10 月 5 日
作者简介	作家，Netswithch 科技管理公司首席运营官，首席安全官，企业安全教父。		
原文发布单位	Netswitch 是一家网络信息技术和服务公司专注于网络安全服务与管理，网络安全与云服务等领域。https://www.linkedin.com/company/130453?trk=vsrp_companies_cluster_name&trkInfo=VSRPsearchId%3A492505037147643379779 7%2CVSRPtargetId%3A130453%2CVSRPcmpt%3Acompanies_cluster		
原文出处	http://www.netswitch.net/the-requirements-for-a-contextual-big-data-behavioral-analytics-engine/		
译者	安天技术公益翻译组	**校对者**	安天技术公益翻译组
免责声明	• 本译文译者为安天实验室工程师，本文系出自个人兴趣在业余时间所译，本文原文来自互联网的公共方式，译者力图忠于所获得之电子版本进行翻译，但受翻译水平和技术水平所限，不能完全保证译文完全与原文含义一致，同时对所获得原文是否存在臆造、或者是否与其原始版本一致未进行可靠性验证和评价。 • 本译文对应原文所有观点亦不受本译文中任何打字、排版、印刷或翻译错误的影响。译者与安天实验室不对译文及原文中包含或引用的信息的真实性、准确性、可靠性、或完整性提供任何明示或暗示的保证。译者与安天实验室亦对原文和译文的任何内容不承担任何责任。翻译本文的行为不代表译者和安天实验室对原文立场持有任何立场和态度。 • 译者与安天实验室均与原作者与原始发布者没有联系，亦未获得相关的版权授权，鉴于译者及安天实验室出于学习参考之目的翻译本文，而无出版、发售译文等任何商业利益意图，因此亦不对任何可能因此导致的版权问题承担责任。 • 本文为安天内部参考文献，主要用于安天实验室内部进行外语和技术学习使用，亦向中国大陆境内的网络安全领域的研究人士进行有限分享。望尊重译者的劳动和意愿，不得以任何方式修改本译文。译者和安天实验室并未授权任何人士和第三方二次分享本译文，因此第三方对本译文的全部或者部分所做的分享、传播、报道、张贴行为，及所带来的后果与译者和安天实验室无关。本译文亦不得用于任何商业目的，基于上述问题产生的法律责任，译者与安天实验室一律不予承担。		

大数据情境行为分析要求

史蒂夫•金，作家、首席运营官、首席安全官、企业安全教父。

当下大数据风靡网络安全领域。我们都希望大数据可以在恶意软件攻击之前阻止它窃取数据、希望智能分析软件可以帮助我们从数千万的数据点中提取最有价值的信息以便发现、阻止网络罪犯窃取重要的数据。幻想软件可以替我们做所有的事情，因此我们不需要很多安全分析员盯着琐细的数据，毫无头绪。

听闻大数据可以办到，但到目前为止进展极为缓慢。

把网络数据一股脑的扔到大数据引擎上是识别、阻止恶意行为的有效途径，这是人们的普遍看法。但这一理论存在两个重大问题：

1.大数据分析工具分析出的内容并没有比数据源提供的更好。

2.无情境分析不能建立威胁关联机制，这对数据安全防护、检测、自动修复毫无用处。

例如像日志文件，网络流量和基线等典型数据源丢失了所有恶意行为的关键指标信息，反而把类似典型数据分析引擎活动记录为良性网络流量。

随着恶意软件不断升级以及业内员工现在大多开启隐身模式操作， 而且他们对这些分析引擎的内部构造也都了解，这使得日志，数据流和基线等数据源越来越难发现恶意软件公认的指示性数据元素。

另外，当今网络协同攻击横跨多级，多个维度。但因检测传统大数据分析引擎的离散事件在情境之外，因此漏掉了一些细微的模式和相关的行为序列，而这些东西当前经常被网络罪犯通过全球威胁地图汇编成有效的深度攻击入侵模型。深度攻击模型是曾经盛行的网络杀伤链模型的简略版，其工作原理是发送载荷，侵占终端、消耗网络、数据渗漏、或破坏信息资产。

因此，为了有效应对深度攻击的威胁，我们必须改变情境数据分析的方法。

一个有效的情境分析引擎除必须分析隐藏恶意行为指示符之外， 还要用恰当的分析类型检测指示符。

同时 ,分析引擎必须在具体威胁邮件情境下使用构建的设计算法检测结构化与非结构化恶意行为。可以通过网外行为模式判断是否是威胁邮件，并发现威胁光谱图。另外，分析引擎必须实时操作数据，在网络入侵和信息被盗之前识别并隔离感染数据。

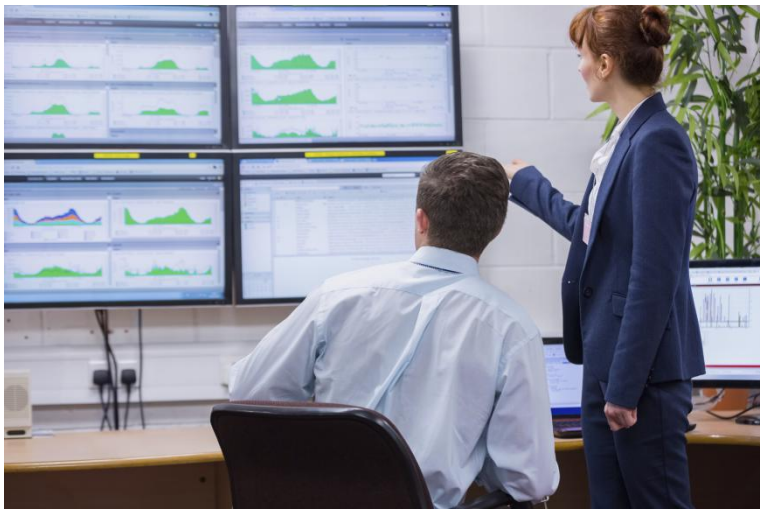
分析引擎核心一贯遵循四种主要推理方法中的一种

演绎推理——演绎推理是基于演绎推论理论，从事物一般规律得出具体的结论。例如，如果 A=B，且 B=C， 那么不管 A， B 包括什么，结论是 A=C。演绎推理是从一般到特殊的推理，其特点是如果最初的前提正确，那结论也一定正确。但演绎推理存在根本缺陷经

常赘述（例如恶意软件总是包含恶意代码）且不受情境输入值的影响，例如，要获得硕士学位，学生必须修满 32 学分。蒂姆拿了 40 学分，所以蒂姆将获得硕士学位，除非他不打算要。

安全分析中，A 大多数时候只等于 B，不过有时候等于 D，所以 A 不总是等于 C。因此使用演绎推理作为检测分析的依据来检验和预测未来，这种方法是有缺陷的。理论上讲，我们不能保证不犯错。

总之，常见的签名系统，例如 IDS/IPS 以及终端安全系统本质上是利用演绎推理。



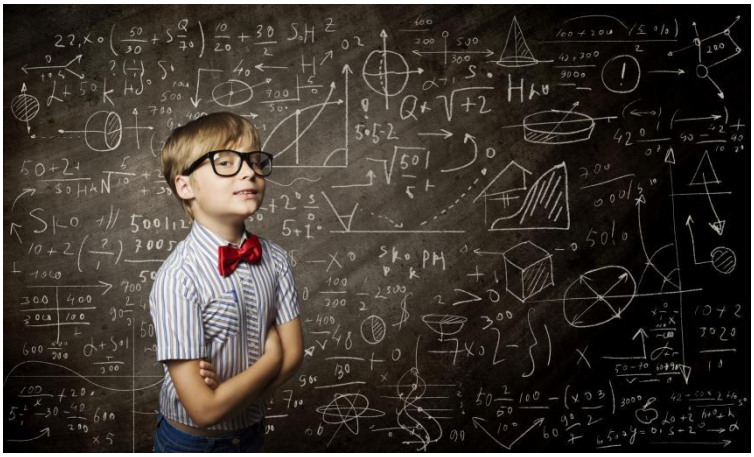
归纳推理——归纳推理与演绎推理正好相反。它通过观察具体的事物得出概括性的结论，即从具体到一般的过程。一般通过细致的观察，理解模型、做总结，然后得出一个解释或理论。

基于归纳推理的分析引擎，其分析结果与概率论相似。即使陈述的所有前提正确，其结论仍可能是错的。举例如下：哈罗德是爷爷。哈罗德秃头了。因此，所有是爷爷的人都是秃头。显然，该结论不符合逻辑。

与演绎推理相比，归纳推理预测未来的方法更优，但显然不完美且产生的结果多变。

一些高级的 IDS/IPS 系统使用启发式归纳推理识别恶意行为。启发法是一种为解决疑难问题提供捷径的规则，被用于帮助监控者在时间有限或信息不足之时做决策。大多数时候，启发式归纳推理会引导你做正确的决策，但对于预防高级威胁来说效果不理想。

启发式归纳推理经常被现在的 IDS/IPS 系统用来基于有限数据输入（例如，已知签名）总结恶意行为发生的概率。



贝叶斯或递归贝叶斯估计推理——该分析方法是不规则导向推理，用于给安全系统提供一般性意见，预测长时间内（例如 30-60 天）将会发生什么。贝叶斯推理是逻辑的分支用于决策和处理概率推论中的推论统计：利用先前知道的信息预测未来的事件。

统计学中，标准离差是用来量化一组数据值变化或离差的计量标准。如果标准偏差接近 0 表示数据点趋于接近该组数据的平均值，而高标准离差则表明数据点更大范围的离散。

大多数基于贝叶斯的安全分析系统，当输出的结果是标准离差正常水平 3，系统会显示异常。贝叶斯推理的目的是通过观察企业设备一段时间内微小波动，建立先验事件库然后识别正常行为模式。该离差结果是基线，作为后续的基准评估未来所有的网络活动或网络行为。

遗憾的是，基线这一概念是有缺陷的会出现极端结果使得无法完全识别威胁。

该方法存在三个严重问题：

1.如果网络或系统在基线创建之前就已经被感染了，那么建立的基线是伪前提。

2.如果内部人员已经在线活动，其活动将停留在表面上且成为“正常”基线的一部分。

3.现在的网络设施和用户行为变得越来越动态，多变，多样涉及到不同的设备和协议，访问方法和入口点，这使得不开启网络禁闭根本无法评估基线。

利用基线作为前提进行贝叶斯推理的分析引擎容易产生大量的数据误报，繁杂且难以优化和管理，要求投入大量的人力且经常遗漏恶意入侵活动。总之，效果不是很理想。

反绎推理——反绎推理是一种来源于对假设观察的逻辑推断形式，它可以解释观察到的现象，寻找最简单、最可能的解释。不像演绎推理和归纳推理，反绎推理的前提不能保证结论正确，这种推理方法更适用于现实世界的恶意网络攻击。

通常，反绎推理始于一组不完整的观察数据，接着寻找最可能的解释。反绎推理通过高效利用手头已有的信息做日常决策，但其信息经常不完整。

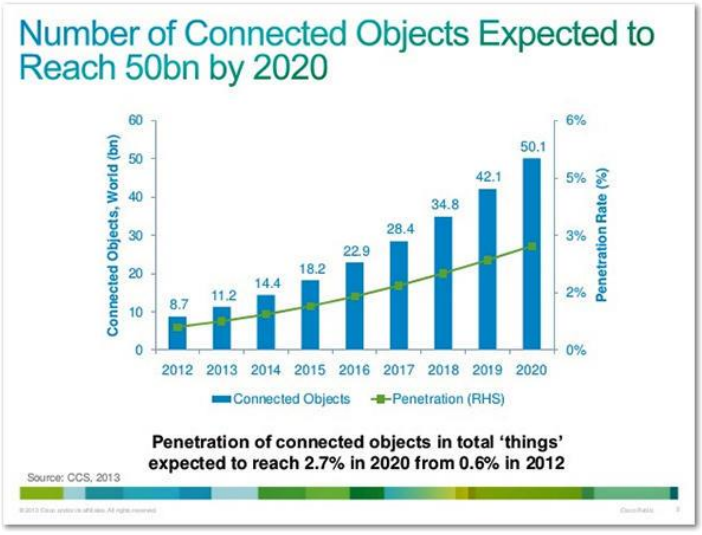
医疗诊断是反绎推理的一项应用：考虑到这些症状，什么样的诊断能够对这一切做最佳解释呢？同样，在法学系统中，当陪审员听取一起刑事案件的证据之后，必须考虑检方或辩方是否能够对所有证据点做最佳解释。虽然证据确凿，该案件可能还存在其它的证据没有得到证实，因此他们会根据已知做最大的推断。

虽然归纳推理要求阐述主体的证据要非常完整。不论是有利还是不利，反绎推理的特征是通过观察一组不完整的数据，要么是证据或者解释，或者两者都有，然后得出最接近预期的结论。

例如，病人可能没有意识到或未能告知其症状导致病情描述不完整或医生无法对某些症状做正确的诊断，但他仍要尽可能做最佳诊断。概率反绎推理是反绎验证的一种形式，在很多领域应用广泛，非常成功。在这些领域是通过获取可能的假设得出结论的，例如通过医疗试验做诊断，审判过程中的推断以及预测恶意软件的存在。

现今大多数安全解决方案专注于事件，试图从数十亿事件中处理数据，发现恶意行为。就其效果来说，该方法极其有限。它非但无法测量，且产生一大堆误报和警报噪音，实际上对解析和减少依赖自动启发分析引擎以便产生可操作信息的数据量毫无作用。连续的事件的确有趣，但单个从情境中拿出来，其提供的有用信息对创建大数据点库收效甚微。

针对网络实体（服务器和客户终端）检查证据和创建感染概要引擎需要什么。这是很重要的区别，因为一个网络可能包含 1000 系统，每一个系统产生 1000 个概要。而像 SIEM 等基于事件系统必须以同样的网络设施从数百万个事件中处理 100 个。但这就给安全分析员带来了三大挑战，极其艰巨：事件过载，误报，缺乏情境。



很多系统利用日志数据企图发现、检测恶意软件，但日志本质是由事件驱动的且现今写得很好的恶意软件不会在日志中留下痕迹。传统的基于日志数据分析的方法要求对数百万个日志事件分类以便把该类事件关联起来，其目的是把数百万不连续的元素转化成行为模式。

现在的数据分析引擎是依据事件而不是系统，因此必须加入完全独立的变量试图构建有意义的行为关系，同时必须同等重视每一个事件。而这要求巨大的数据处理能力（成本高昂）且会不可避免的产生前面我们已经提到的问题——一定比例的低信号噪音，很多误报。更糟的是出现漏报。

事实上，人们对贝叶斯递归推理，演绎推理和归纳推理引擎越来越担忧是由于出现了漏报，这表明感染系统修复后依然会出现问题。为了弥补这一趋势，引擎的灵敏度经常调整的过于精确而犯错——从而产生更多的漏报。

就像我们之前描述的那样，大海捞针只是成功解决问题的开端。一套有效的行为分析恶意软件解决方案还要求有情境分析。没有内容的网络分析会产生很多噪音，缺乏可操作数据。

我们需要做的是限制恶意活动所带来的风险内容。例如，发现某个漏洞正在对特定的系统进行攻击，其带来的价值有限。而发现漏洞正在对缺少补丁的系统攻击，使得系统在漏洞面前很脆弱，这对漏洞攻击具有极高的价值。

情境是指系统情境和情境迹象的实时本质。有效分析必须追踪实时活动以便成功识别、确定真正的恶意行为。仅靠依赖沙箱引爆潜在状态下的恶意载荷和二进制配置的安全解决方案显然在上一个系统扫描之后就过时了。另外，恶意攻击媒介实时变化，不会当其它媒介正被引爆分析时静止不动。

成功的行为分析引擎会将实时漏洞评估与正在进行的威胁活动相关联，这涉及到引入完整性度量和验证（IMV）扫描和 LDAP 实时给安全分析员决策发送必要的情境(例如，活动目录）。

要想为归因威胁信息在安全社区交流，需通过连接器，入站 REST APIs、行业标准符号

与外部供应商产品实现完整的互操作性。要获得其它的情境，需要引擎通过蜜罐技术，技术合作方和基于行业组织（例如，NIST,MITRE,US-CERT）出版的安全公告整合日常获取的威胁情报。

另一种方法是在网络中横向移动把南北向变成东西向。具体来说，我们需要监测机制就位发现恶意内部行为和分析识别其真实行为，恶意以及操作了什么内容。

大多数工具试图识别恶意内部人员的方法是使用 NetFlow 数据。NetFlow 原本是 Cisco 开发的网络协议用于帮助网络工程师规划网络设施、优化性能和更好的管理流量和路径。但作为分析数据源，NetFlow 的效果有限。

NetFlow 从没有打算做安全工具，一部分原因是信息有限，但也是因为包含的所有信息本质是历史信息，且没有实时面流入，因此恶意内部人员操作必须实时了解信息。

为了解决 NetFlow 存在的限制，一套有效的行为分析引擎必须使用数据交换协议实时操作并创建本质上稳定的数据流。同时需要协议在数据流动之时建立初始记录然后生成实时更新记录。

建立情境，记录需要包含基本流指标之外的特性，包括网络地址，服务端口、地理位置指示器、数据计数器、连接状态，威胁标签、DNS 交易和连接信号等（超时，重置）等。

这将是以网络访问管理为中心到以流熵管理为中心的模式转移，它要求内部完整操作、跨域网络活动和数据转移能够被检测。然后才可能发现数据横向移动，发信号（回调函数，beacon 工具，背景连线通讯）和使用基于流逻辑事件实时关联数据渗漏。

这将为战术参与干预活动流打开了大门危害甚大。基于政策威胁协议可以使用一组变量、标准和请求行为进行定义，使得其可扩展、可定制。政策变量可以提供关于监测系统的属性（轴点），标准可以识别基于流变量的预选节，而请求行为可以规定自动修复和事件响应。

另外，一套默认的全球政策应该可以获得解决方案用于监测危险行为（可疑），而一款简单的 API 应授予安全分析员基于内联网拓扑结构和存储的详细信息定义当地网络政策。此外，政策也应可以容易转移（根据内联网网络拓扑结构隐私条款）以便在安全社区内分享威胁定义。

当然，我们需要通过把数据转换成集中可用的指示器集以转化关联多种外部威胁源。安全分析员利用这些指示器预判当地 IT 设施中潜藏的更广泛的威胁。

为了有效反击、打击恶意软件和恶意内部人员，我们需要对数据分析优化而不是大数据分析。为了深层次分析，我们需要正确的分析方法，使用合适的检测引擎、实时发送情境内保护的系统信息。