第八章　数字取证

数字取证

- **概述**

- 数字证据从无到有、发展成为一种超越传统的全新证据形式在各种案件中出现，用于发现、寻找数字证据的数字取证技术越来越为人所重视，尤其是司法部门。
- 进入21世纪后，在大数据时代背景下，数据挖掘技术将是用于数字取证的合适的科学方法。

目录

- 数字取证技术
  - 数字取证的定义
  - 数字取证的发展
  - 数字取证的原则、流程、内容和技术
  - 数字取证面临的挑战
- 数据挖掘在数字取证中的应用
  - 文献概览
  - 现有用于数字取证的数据挖掘技术和工具
  - 电子邮件挖掘
  - 数据碎片分类
  - 文档聚类

数字取证技术

1.数字取证的定义

2.数字取证的发展

3.数字取证的原则、流程、内容和技术

4.数字取证面临的挑战

- 计算机相关案件的不断出现，使得计算机证据逐渐成为新的诉讼证据之一。
- 计算机证据对司法界和计算机安全科学领域提出了新的挑战。因此作为计算机领域和法学领域的一门交叉学科——计算机取证(Computer Forensics) 成为人们研究与关注的焦点。
- 随着信息技术的发展，设备不再局限于计算机，而是各种数字设备，计算机取证这一概念也被数字取证（Digital Forensics）所替代。

数字取证的定义

- **数字证据**

传输于计算机系统或网络间、存储在数字设备或介质中，和案事件事实相关、有证据价值的数据。

特点：

1. 容易被改变或删除，并且改变后不容易被发觉
2. 多种格式的存储方式
3. 易损毁性
4. 高科技性
5. 传输过程中通常和其他无关信息共享信道

数字取证的定义

- 数字取证就是以便于、促进重构犯罪事件，或帮助预见未授权的破坏性行动为目的，使用科学衍生并已被证明的方法保存、收集、确认、识别、分析、解释、记录和展现从不同数据源获得的数字证据的活动。----2001年第一届数字取证研究国际会议（DFRWS）技术委员会
- 数字取证是揭示和解释电子数据的过程。其目标是通过收集、识别、验证数字信息开展结构化调查的同时保全证据最原始的形式，以重构过去事件。----技术百科（Definition from Techopedia）
- 。。。。。。

数字取证的发展

- **境外数字取证发展情况**
- **境内电子物证检验鉴定发展情况**

数字取证是伴随着计算机犯罪事件的出现而发展起来的。目前，国内有关数字取证方面的研究和实践尚处初步阶段，与西方先进国家尚存较大差距。

境外数字取证发展情况

- 美国是最早研究电子数据检验的国家，早在1984年美国FBI实验室和其他执法机构就开始研究电子数据检验的程序和方法。
- 1992年-- FBI建立了计算机检验和响应机构
- 至今-- FBI下属50几个办事处均有CART实验室及人员，每名人员除了具有专业基础知识外，还必须经过FBI组织的七周以上的专门培训，每年还需要3周时间进行电子技术培训和参加有关学术会议，以更新知识。美国至少有70%的法律部门拥有自己的计算机取证实验室

境外数字取证发展情况

- 英国在FSS中专门设置开展了电子数据检验的部门，主要工作任务是快速目标搜寻、计算机检验、移动电话和掌上电脑检验、手机基站信息分析、视听鉴定检验等。
- 香港警务督察处刑事部于1993年在商业犯罪调查科设立了电犯罪组（电脑鉴证组），负责电子数据检验工作，内容包括：数据恢复、互联网取证检验、服务器日志检验、密码破译、手提电话、邮件、存储介质检验等。

境内电子物证检验鉴定发展情况

- 在执法部门中，以公安机关的电子数据取证机构发展得最为完善、业务能力最强。在公安机关内部，各业务警种都或多或少配备了电子数据取证设备，培养了取证人才以应对各自领域中的电子数据取证需求。

数字取证的原则、流程、内容和技术

- 数字取证原则
- 数字取证流程
- 数字取证内容、技术

数字取证原则

现场保护，快速收集原则

可靠性原则

可复现性原则

完整性原则

全程记录原则

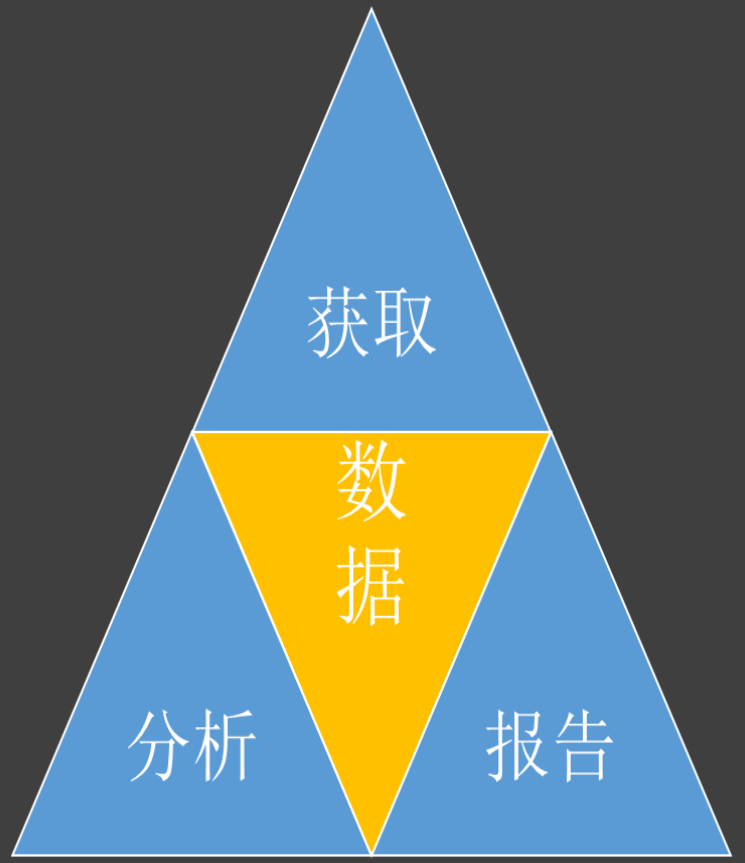
数字取证原则

- **应对大数据的挑战，需要扩展以上原则**

- 工具和方法的验证和可靠性在大数据场景的情况下获得更大相关性
- 可复现性作为数字取证基本信条很可能要被迫放弃
- 对于文档记录，发布关于所使用的工具和方法的证实结果的数据

数字取证流程

数字取证主要包括三个核心流程：数据获取、数据分析、数据报告



数字取证流程

也可划分为5个步骤

- 识别：判定事件/案件类别
- 保全：开始证据保管链，扣押数据和文档
- 收集：镜像存储设备、恢复数据
- 分析：检验所恢复数据，搜寻潜在证据
- 报告：记录事实和发现，总结证据，准备证言

数字取证流程

根据唯一可用的国际标准ISO/IEC 27037，我们在这将其描述为

识别	• 搜索、识别和记录可能包含数字证据的现场物理设备。重点在于识别、定位潜在证据，并注意其可能位于非常规位置。
收集	• 前一阶段识别的设备可被收集、转移到分析场所或现场进行证据获取。
获取	• 涉及对潜在证据源制作镜像，最理想的是保持与数据源完全一致。
保存	• 证据完整性，包括物理和逻辑两方面，必须全程得到保证。
分析	• 诠释所获取的证据数据。它通常依赖于案情、调查目标或焦点。
报告	• 将数字调查结果向相关方传达

针对大数据场景，取证流程可能需要作出调整

识别和收集：这一阶段的挑战是在现场及时选择证据。

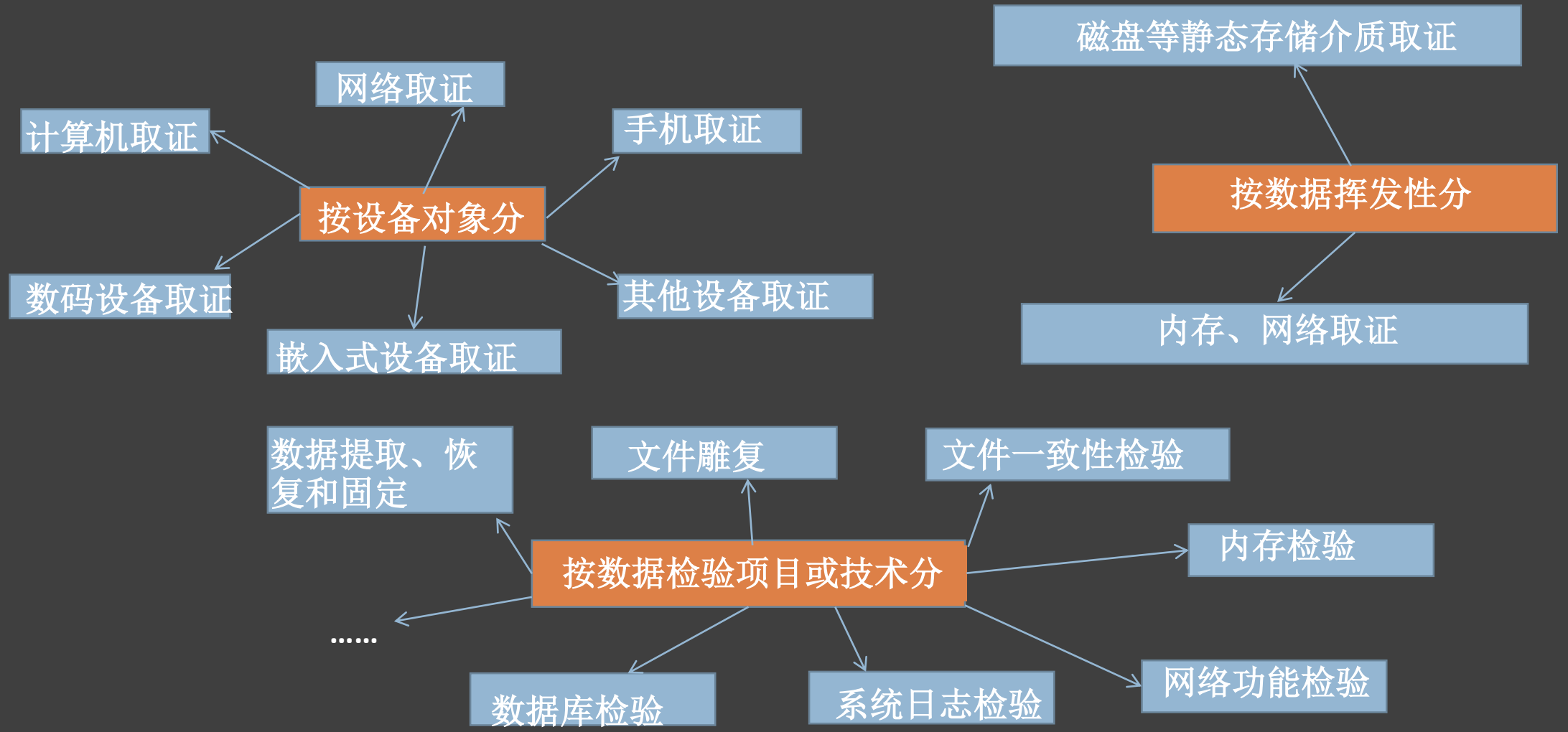
获取：当经典的按位镜像由于检材容量问题不再切实可行时，优先级区分程序或分类工作应被开展，并恰当地予以说明和记录

保存：遵循法律要求安全保存所有检材

分析：集成源自数据科学的方法和工具意味着对当今盛行的“香肠工厂”式的取证（不熟练的操作者严重依赖点击大集成工具实现分析任务）的超越。

报告：对于使用数据科学概念开展分析的最终报告应包含对所使用工具、方法的正确评价

数字取证内容


按设备对象分
网络取证
计算机取证
手机取证
数码设备取证
其他设备取证
嵌入式设备取证
按数据挥发性分
磁盘等静态存储介质取证
内存、网络取证
按数据检验项目或技术分
数据提取、恢复和固定
文件雕复
文件一致性检验
内存检验
……
数据库检验
系统日志检验
网络功能检验

数字取证技术

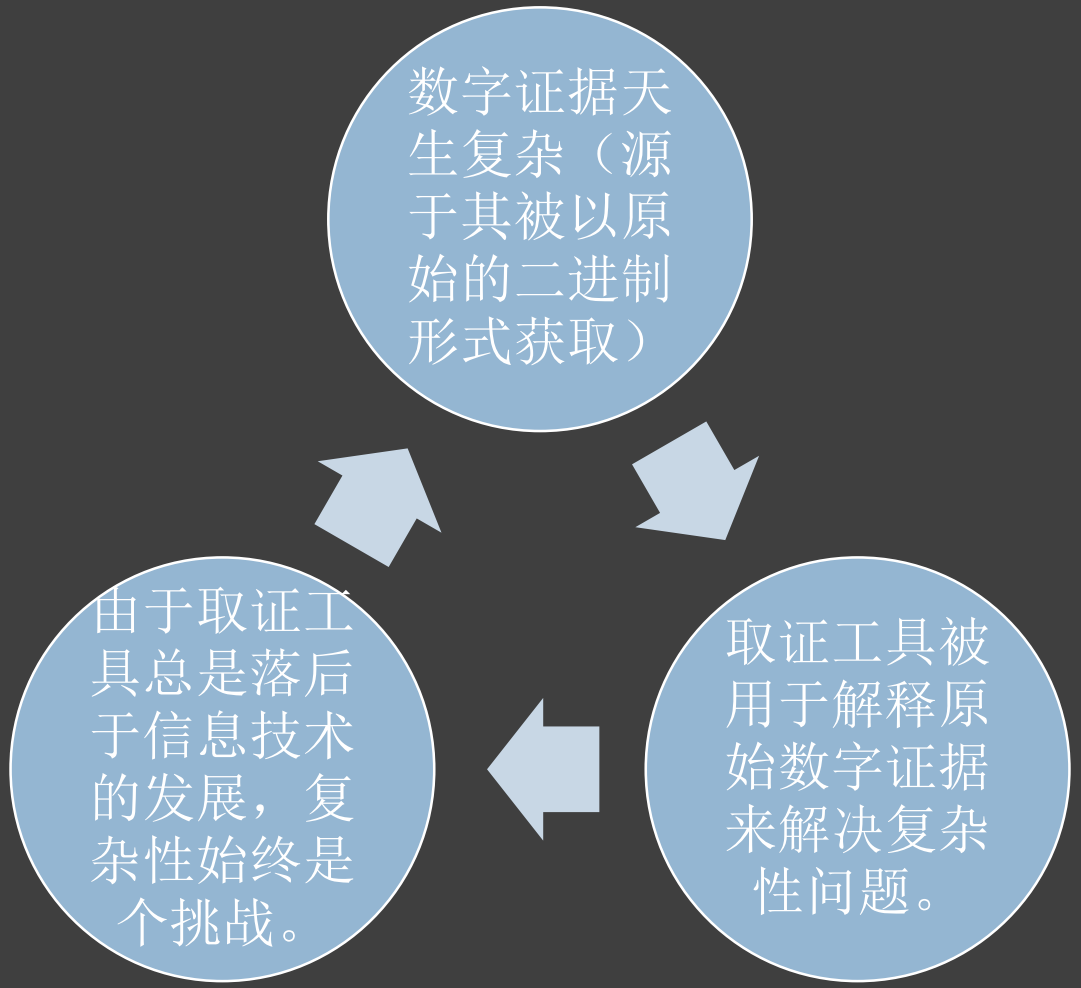
磁盘镜像，数据哈希验证，存储介质的只读访问，文件签名校验，数据恢复和雕复，信息提取、过滤和搜索（包括正则表达式），数据加密和解密，信息隐藏及显现，数据解析和展现，代码反向工程，数据库、数据仓库技术，并行、分布式处理，虚拟仿真等等

数字取证面临的挑战

- 复杂性
- 多样性
- 数据规模
- 加密和云计算
- 一致性和相关性问题
- 统一时间轴问题
- 现实与理想的差距

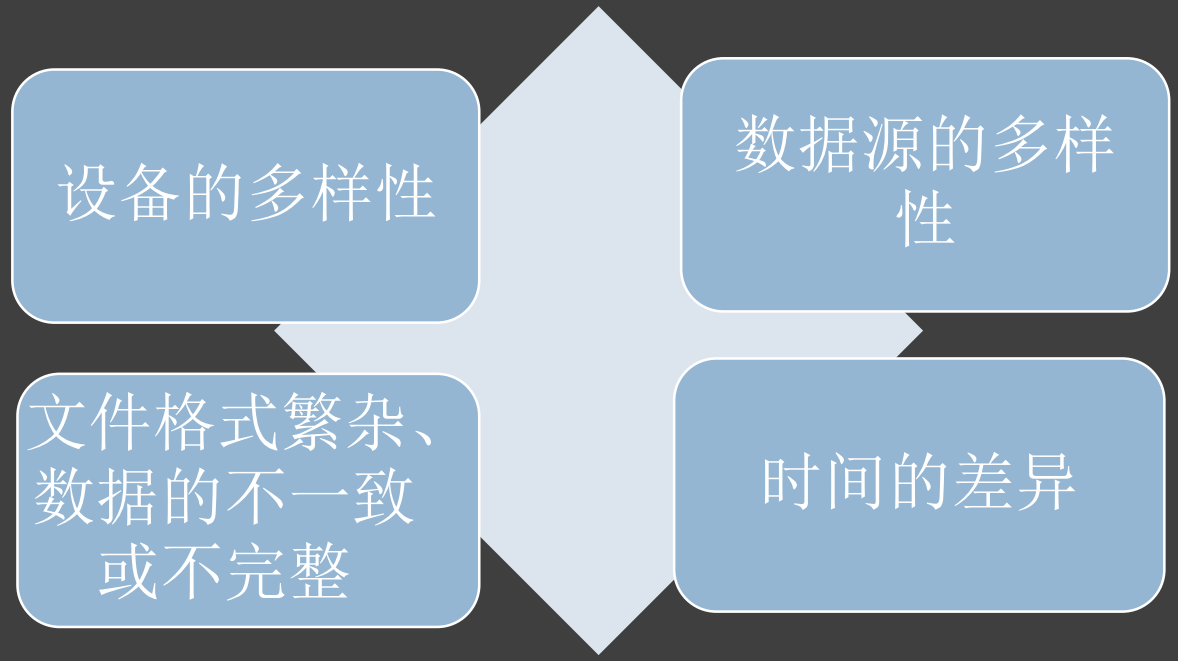
2005年以后，随着云计算、物联网的出现，人类社会已经处于“到处是终端，无处不计算”的数字环境里，数字取证更是面临着诸多挑战

复杂性


数字证据天生复杂（源于其被以原始的二进制形式获取）
取证工具被用于解释原始数字证据来解决复杂性问题。
由于取证工具总是落后于信息技术的发展，复杂性始终是个挑战。

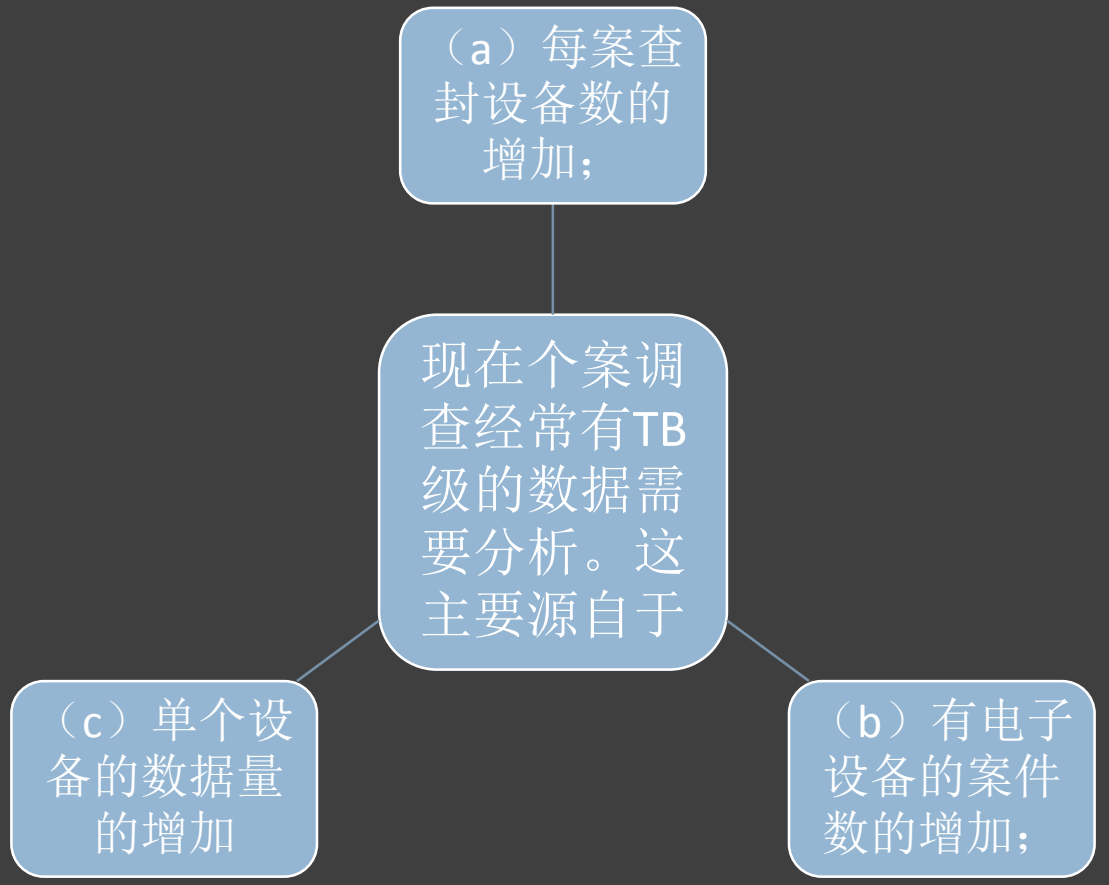
多样性

多样性是数字取证的根本性挑战。



面对多样性，目前还没有能适应所有种类电子数据的检验分析的标准方法。

数据规模


（a）每案查封设备数的增加；
现在个案调查经常有TB级的数据需要分析。这主要源自于
（c）单个设备的数据量的增加
（b）有电子设备的案件数的增加；

数据规模

首当其冲的是单一案件检材的平均容量呈爆炸性增长。诱因有：

- 硬盘和固态存储价格显著下降，以至于单台计算机或存储设备的存储容量上升。
- 磁存储密度提升和固态可移动介质（U盘，SD和其它存储卡）的大量采用。
- 在世界范围内，智能手机、平板电脑等个人移动设备迅速普及。
- 虚拟化技术和现代多核处理器使得云服务在个人和商业领域的推出和持续采用成为可能。
- 网络流量更加成为案件证据来源，且其数据量在近十年一再飙升，不管是宽带还是3G-4G移动网络。
- 连通性无处不在，随着不久将来实现向IPv6过渡，“物联网”越来越近。

数据规模 -不断增大的数据积压带来严重影响。

- Vidas等在2014年指出，不断增加的数据容量导致案件在取证实验室的积压，而许多案件需要几天或几周的分析时间与其需要等到快速分析的期望相去甚远。
- Noel and Peterson (2014)讨论到，大数据问题，正在导致数字调查复杂化，决策低下，机会丧失，证据难以发现和生命的潜在丧失。

加密和云计算

- 云存储数据可以是分布在不同地点的大量数据，甚至是同一数据被碎片化存储在不同位置，给取证分析带来许多困难。
- 加密还使得数据即使能被恢复，也难以被处理分析。

一致性和相关性问题

- 数字设备、数据的多样性，加之数据规模的膨胀导致需要分许多个设备的案件越来越多，使得当多种来源的数字证据被识别、调查时，不仅分析它们非常重要，证实、关联这些不同来源的数据以说明一致性也必不可少。

统一时间轴问题

- 不同时区参考和时间戳解释
- 时钟偏移/漂移等问题
- 在生成统一时间表时的语法方面问题

生成统一的时间轴面临

现实与理想的差距

- “犯罪现场调查效应”导致不切实际的期望。
- 数字取证对人员、设备要求高。
- 设备和培训跟不上。
- 司法层面：比如司法管辖权问题、标准化的国际公认法律缺乏、法官的数字取证技术知识缺乏等问题

数据挖掘在数字取证中的应用

- **研究成果概览**
- **现有用于数字取证的数据挖掘技术和工具**
- **电子邮件挖掘**
- **数据碎片分类**

- 数据挖掘是一个交叉的研究领域，它采用统计学模型、数学方法、机器学习等方法，可以发现大型数据集中信息的未知模式和关系
- 将数据挖掘技术于电子取证，以解决信息容量不断增长的问题，是非常重要的数据挖掘技术应用。

研究成果概览

- 数字挖掘技术在数字取证的以下一些方面的应用：
- 1.入侵检测
- 哥伦比亚大学的Stolfo等已经构想出一种基于审计源数据挖掘实现的入侵检测系统。基于该模型，系统得以发现大量的攻击特征和正常模式，用于构建动态可配置的模型组

研究成果概览

- 2.图像挖掘
- 昆士兰大学研究人员与澳大利亚防卫科技组织合作利用数据挖掘技术设计出一个图像挖掘系统。“该系统能被一个分级支持向量机训练以检测那些在空间或非空间限制下部件组成的对象和场景”。

研究成果概览

- 3.犯罪网络分析
- 在一个叫做COPLINK的美国国家科学基金（NSF）资助的数字管控计划项目里，研究人员利用数据挖掘技术在执法过程分析数据。
- 这种分析涉及四个步骤：网络提取、下属组织检测、互动行话发现和核心成员识别
- 2003年，美国亚利桑那州大学人工智能实验室介绍了COPLINK项目的案例研究概况

研究成果概览

- 4.Email内容挖掘
- 目前已经存在用于垃圾邮件检测、控制和邮件自动归档等各种任务的email内容分析的相关研究。
- 纽约哥伦比亚大学研究人员开发了一款邮件挖掘工具EMT，用于帮助执法部门和数字取证专业人员分析邮件并使其成为呈堂证据。

研究成果概览

- 5.严重性罪犯行为挖掘
- 英国伍尔弗汉普敦大学的研究者和伯明翰警察局，应用数据挖掘技术将严重性犯罪予以关联。他们采用Self Organizing Maps (SOM)——一种亚型人工神经网络用于这种分析。分析人员在与该案件呈现出强相似性的簇类中判定罪行

研究成果概览

- 6.在取证数据中关联/分类/聚类/预报，可视化展现
- 关联—识别相互关系
- 分类--基于相似性将数据划归相应组别
- 聚类--定位各组潜在事实
- 预报--发现能导致有用推断的模式
- 可视化--对数据进行可视化展现[

研究成果概览

- 7.Map-Reduce
- 一个大型并行任务框架，当数据集没有包含许多内部相互关系时运作良好。
- 对于类似文件碎片分类这样的任务是适于在Map-Reduce范式中模型化处理的。判断来自文件系统镜像或未分配空间的文件碎片归属于哪一个特定文件类型在数字取证中很常见。

研究成果概览

- 8.内容提取技术
- 内容提取技术通常被用于文本数据、多媒体数据、互联网数据、空间数据、时间系列或序列数据以及复杂数据
- 数字取证数据可能由来自各种数据源的结构化、半结构化数据，噪音和正常数据等组成
- Shannon (2004)勾勒了一种称作Forensic Relative Strength Scoring (FRSS) 取证相关强度评分的内容挖掘技术
- Noel and Peterson (2014) 提出使用隐狄利克雷分布Latent Dirichlet Allocation (LDA)自然语言处理方法去处理数字取证数据中的用户数据。

研究成果概览

- 2005年，HP公司运用数据挖掘技术解决在巨大文档库中查找相似文件的问题。
- 2006年，Galloway和 Simoff 通过案例研究试验重新定义网络数据挖掘方法。
- 2007年，Beebe和Clark在工作中提出数字取证文本字符串搜索结果的检索前和检索后聚类。

现有用于数字取证的数据挖掘技术和工具

数字取证任务	数据挖掘技术	工具
数据恢复, 数据生成和预处理	统计测试分析, Bartlett球体检验, Kaiser-Meyer-Olkin (KMO)统计量检验	Recuva, FTK, Encase, Sleuth kit/Autopsy, ProDiscover
数据分析	聚类 – K均值, 最大期望算法，分级聚类	weka
	分类 –监督学习- 决策树, 神经网络,支持向量机, 朴素贝叶斯	weka
	非监督学习 – 主成分分析法PCA， Karnohuen映射	-
	频繁模式挖掘/关联规则挖掘 -Apriori, Eclat	weka
	实体识别	Lingepipe
	可视化	Cyber Forensics Time Lab
	统计分析和异常检测	EMT/MET
	递归数据挖掘	-
	网络仿冒（钓鱼）	Invisible Witness
	回归	-

电子邮件挖掘

- **email和文本挖掘在邮件数据特性上的分界**

- Email在邮件头包含可被开发利用与各种email挖掘任务的额外信息。
- Email中的文本特别短，通常比故事、用户手册等许多文档来得简洁、短很多。
- Email经常被马虎书写，因此不能保证语言被良好组织。
- Email是个人的，通用技术难以有效应对个体。
- Email是针对特定用户的数据流，其信息针对的目标阶层的概念和分布也许随着时间而变化，相对于那个用户收到的信息。
- Email很可能有噪音。
- 由于隐私问题，想要有公开的email数据用于实验，相当困难。

电子邮件挖掘

- **电子邮件挖掘主题**
- **Email挖掘所使用的算法**

- 邮件挖掘被不同研究人员开展用于从email中提取不同信息。

电子邮件挖掘主题

1.作者身份属性分析（联系人分析）

- 作者身份归集意味着从一群潜在嫌疑人中识别出匿名邮件最可能真实的作者。
- 每个人都有独特的身份、特点和写作风格。

电子邮件挖掘主题

2. 内容分析

- 内容分析是依赖科学方法（包括关注到客观性、互为主体性、先验设计、可靠性、有效性、普适性、可复制性和假设检验）对信息的总结、量化分析，且不局限于所度量的变量类型或消息被创建和呈现的上下文。
- 内容分析可被用于实现电子邮件自动回复、电子邮件间关系分析、邮件分类等容分析。

电子邮件挖掘主题

3.钓鱼

- 钓鱼是一种欺诈，email用户被诱骗提交被用于身份盗用的个人信息。
- 它是互联网上增长最快的欺诈行为。
- 钓鱼邮件结构特征包括文本内容、邮件内容、脚本、表格、图片或徽标，超链接，表标签，伪造标签等。

电子邮件挖掘主题

4.垃圾邮件过滤

- Yong Hu等推荐模糊分类方法区分垃圾和正常邮件。他们提出的垃圾过滤器包好四个部件，即“特征提取器”、“模糊分类算法”、“打标签算法”、“调整算法”。
- Chun Wei等集中注意力在垃圾邮件的高级分析上，通过考虑邮件信息的11种属性：邮件ID，发送者IP地址，发送者等
- Salvatore J. Stolfo等给出一个名叫EMT（Email挖掘工具）的数据挖掘系统，用于核心安全应用，以检测病毒传播，“垃圾邮件程序”活动和安全策略违反。

电子邮件挖掘主题

5.邮件分类（邮件归档）

6. 邮件网络属性分析

- 即分析email网络的关键属性，比如网络总体结构、关系强度和组织结构。

7.邮件可视化

- 利用可视化技术帮助用户识别、提取和总结隐藏在大量邮件中的有用信息。

Email挖掘所使用的算法

- 用于邮件挖掘的有各种不同算法，包括支持向量机算法[41][42]、朴素贝叶斯算法、最大期望分类算法、K-均值分类算法、二分K-均值分类算法、凝聚分类算法、基于行为模型等等。对于邮件挖掘的不同主题任务，所使用算法也有区别，比如垃圾邮件过滤偏向于利用分类算法，邮件分类偏向于利用聚类算法，邮件网络属性分析偏向于使用关联规则挖掘算法。

数据碎片分类

数据碎片分类是数字取证的一个重要内容。重构犯罪活动经常需要对已删除、被隐藏或未分配的数据进行分析。因此从数据碎片中重构证据的需求稳步增长。

- 数据碎片的来源各式各样，在各种存储介质上都能发现残留的数字数据，比如硬盘、优盘、内存转储文件或计算机网络数据包。
- 数据碎片分类的典型应用是诸如文件雕复、内存或网络数据转储分析等

Rainer Poisel等的数据碎片分类研究

- Rainer Poisel等2013年调查了超过100篇反映数据雕复领域最新发展水平的研究论文，将数据碎片分类技术分成如下主要类别：基于签名的方法，统计方法，计算智能的方法（人工智能方法），虑及上下文（或称内容感知）的方法，和其它方法。

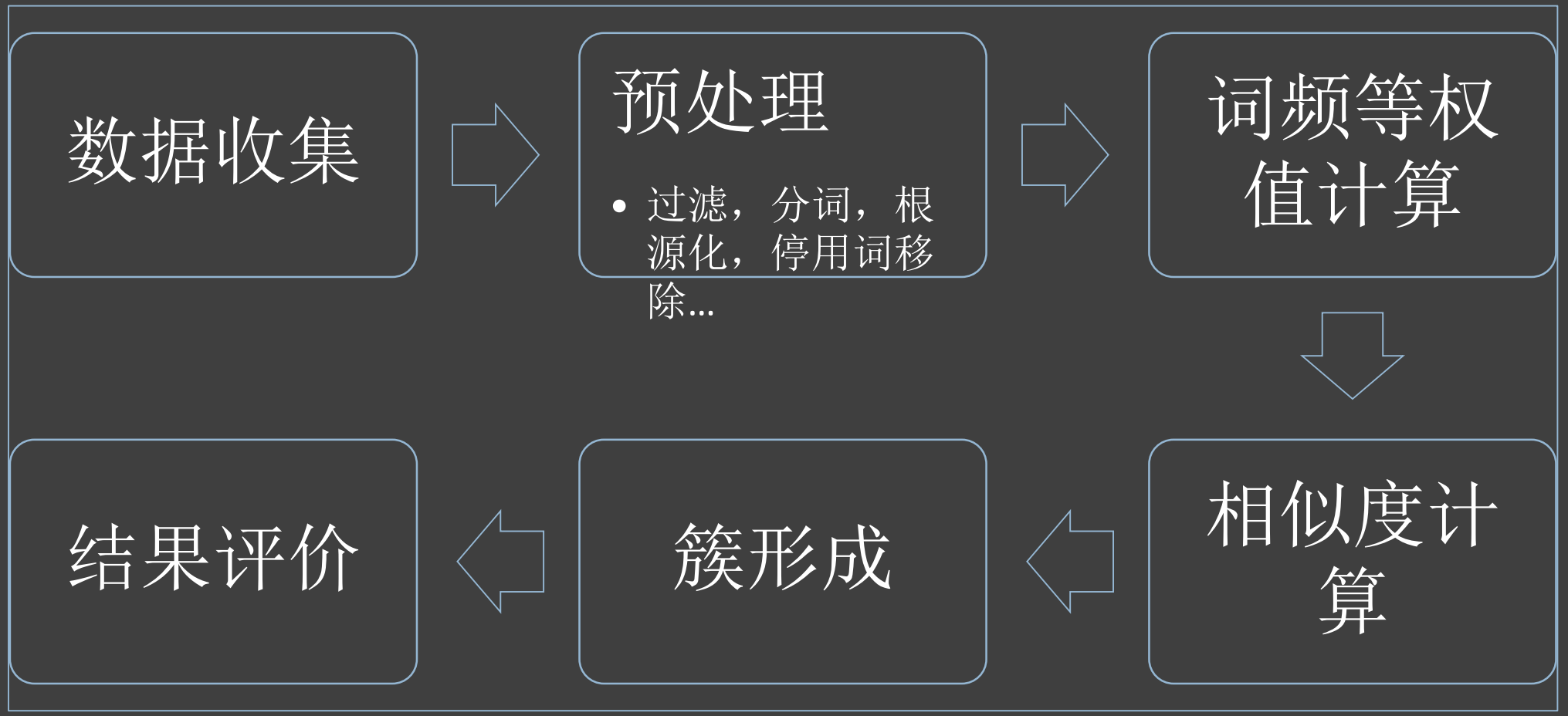
Ahmed、李等的数据碎片分类研究

- Ahmed等[45][46]分析了具有高频率字节模式（1-gram特性）的流行的分类器（比如，神经网络，kNN，SVMs等）的准确率。
- 李等[48]详细说明使用支持向量机的文件类型识别。

文档聚类

数字取证分析经常涉及大量文件的检验。在所有这些文件中，那些检验人员感兴趣的文件需要被快速找到。现有用于分析文档集的数字取证工具提供多层级的搜索技术用于响应问题、生成调查相关的数字证据，但是缺乏允许调查人员按照感兴趣的某一特定主题寻找文档或基于给定主题对文档集分组的功能。文档聚类是将相似文档归入同一类，以利于有效提取信息、减少查找时间和空间、移除异常点、处理高维度数据、提供相似文档摘要的过程。在数字取证调查中，文档聚类的结果是提供不同类别的文档，让取证人员可从中只分析与所调查案件相关的部分。

文档聚类流程



文档聚类的挑战

- 1.选择合适的用于聚类的文档特性。
- 2.选择合适的文档间相似性度量。
- 3.选择合适的能利用前述相似性度量的聚类方法。
- 4.以有效方式实现聚类算法，使之与所需内存、CPU资源一致可行。
- 5.寻找方法以评估所开展的聚类质量。

文档聚类算法

- 利用算法进行文档聚类的主要目的是促进从被分析文档中发现新的、有用的知识[55]。聚类算法需满足的主要要求有：可伸缩性，能够处理不同类型属性，强抗噪性，高维性，对输入顺序不敏感性，可解释性和可用性等[55]。用于文档聚类的算法有：K均值、K中心点等划分算法，单链、全链、平均链等层次聚类算法，基于簇相似的划分算法（Cluster-based Similarity Partitioning Algorithm，CSPA），模糊C-均值(fuzzy C-means,简称FCM)算法。表8.3列出了部分文档聚类相关文献所涉及的聚类算法[56]。

本章小结

- 使用数据挖掘技术辅助数字取证至少能带来三重目标：1）减少系统和人力处理时间；2）改进数据分析的效率和质量；3）减少成本。
- 另一方面，使用数据挖掘技术辅助数字取证也存在潜在限制。首先，这些技术在数字取证专业未得到足够的实际测试。其二，数字取证或数字调查团体对数据挖掘技术缺乏了解。
- 数据挖掘技术在数字取证领域的应用尚处婴儿期，在数据挖掘技术被成功应用并弥漫至整个数字取证、调查团体之前，还有许多工作要做。推动这些技术的建议有：提升对数据挖掘技术的意识和了解，培训数字调查人员使用这些技术，创建在数字取证调查中使用这些技术的框架，活跃将数据挖掘应用到数字取证和调查领域的研究。