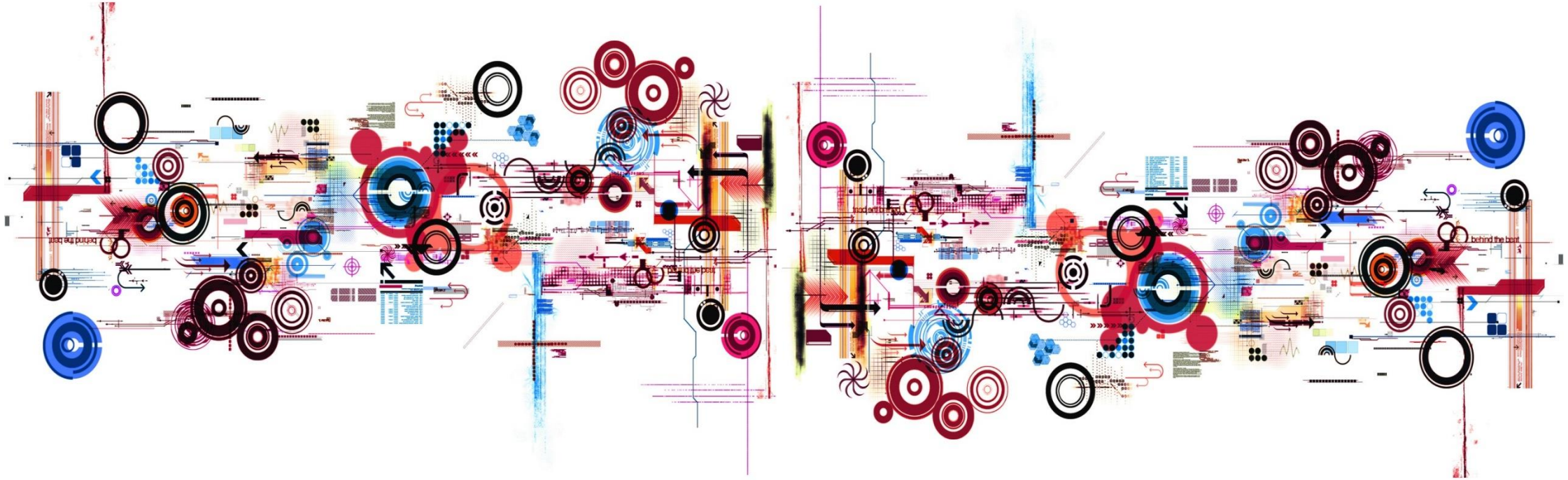
eDiscovery & Dokumentenreview

Technische Möglichkeiten beim Einsatz von eDiscovery-Werkzeugen



Die FAST-DETECT GmbH

Deutschlands größtes Sachverständigenbüro für IT-Forensik

- Gegründet 2003
- Beauftragt von Anwaltskanzleien und Sicherheitsbehörden in ganz Deutschland und Österreich
- Renommierte Kunden aus der Privatwirtschaft
- Deutlich über 5.000 Gutachten (viel Gerichtserfahrung)
- ISO 9001 und ISO 27001 Zertifizierung
- Vier mal in Folge ausgezeichnet als „*Great Place to Work*“
- 36 regelmäßig polizeilich überprüfte Mitarbeiter

Ihr heutiger Referent

Fabian Unucka, 1981
Diplom-Informatiker (Univ.)

Sachverständiger für IT-Forensik
IT-Leiter

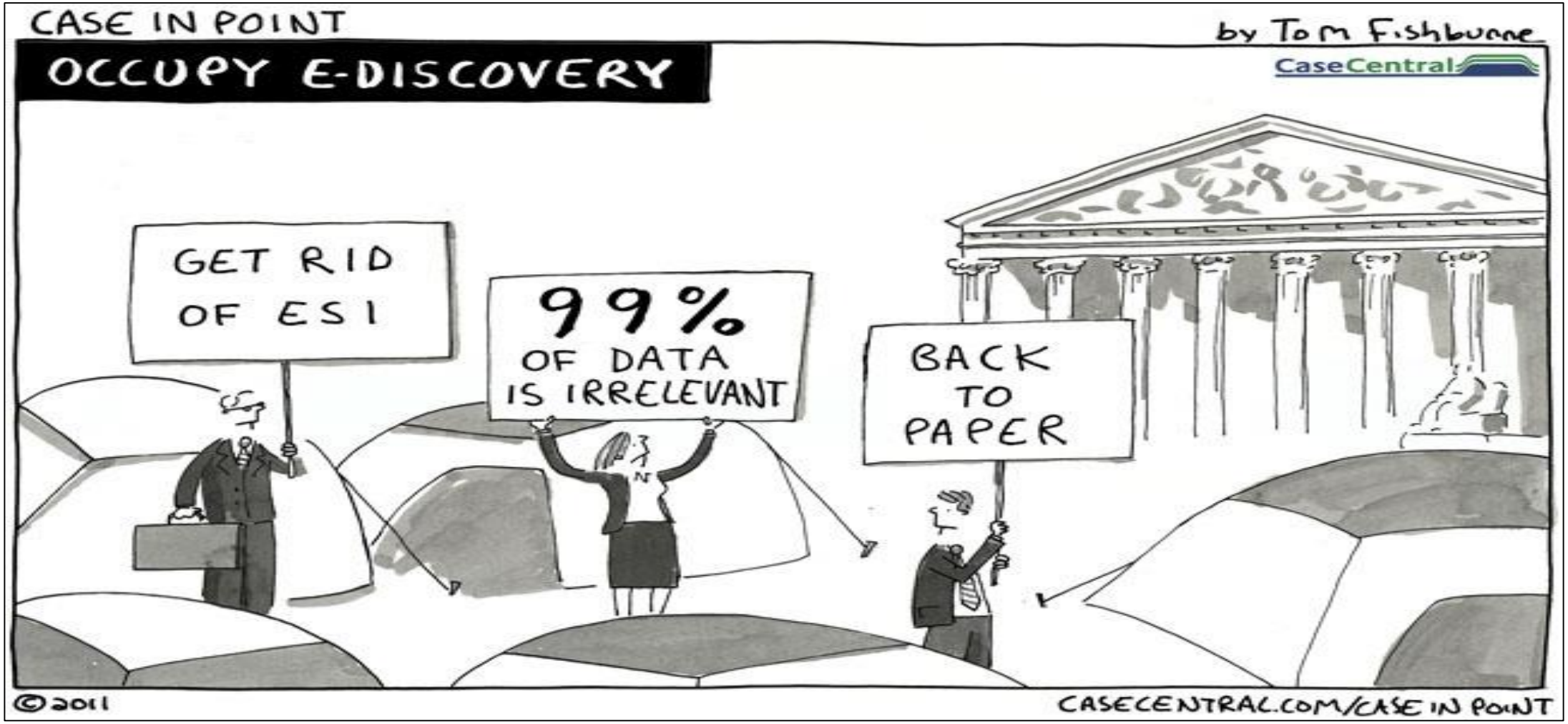
Das Datenvolumenproblem

Ausgangslage

- Eine Kartellermittlung
- 120 Mitarbeiter stehen im Fokus relevante Informationen direkt oder indirekt zu besitzen
- Das gesamte Datenvolumen der 120 Mitarbeiter inkl. der Shares, auf die sie Zugriff hatten, summiert sich auf ca. 160 Terabyte auf (ca. 16 externe Festplatten)
- Zeitrahmen: 3 Monate

Frage: *Wie?*

Das Problem der Dokumentenreview



eDiscovery im amerikanischen Recht

- Im Jahr 1938 wurde im amerikanischen Recht durch das Federal Rules of Civil Procedure (FRCP) die Form des evidence discovery in Gerichtsverfahren und zivilen Streitverfahren eingeführt. Während dieser Phase (Discovery) kann die Gegenparty jegliche Informationen, die relevant für den Fall sein können, anfragen. Die andere Partei muss diese Anfrage beantworten oder beweisen können, wieso diese Informationen nicht geliefert werden können.

- Im Jahr 2006 wurde eine Novellierung des FRCP verabschiedet. Nun umfasst die Discovery Phase jegliche electronically stored information (ESI).

Begriff: Electronical Stored Information (ESI)

Definition

Electronically stored information (ESI), for the purpose of the Federal Rules of Civil Procedure (FRCP) is information created, manipulated, communicated, stored, and best utilized in digital form, requiring the use of computer hardware and software.*

Datenvolumen ESI:

Jeder Mitarbeiter erzeugt im Durchschnitt 3.000 MB an ESI pro Jahr.**

* Electronically Stored Information: The December 2006 Amendments to the Federal Rules of Civil Procedure, Kenneth J. Withers, Northwestern Journal of Technology and Intellectual Property, Vol.4 (2), 171

** Stand 2018

Mit dem richtigen Tool zum schnellen Erfolg?

opentext™

Vound

nuix

Das ‚richtige' Werkzeug kann eine hilfreiche Unterstützung bieten.

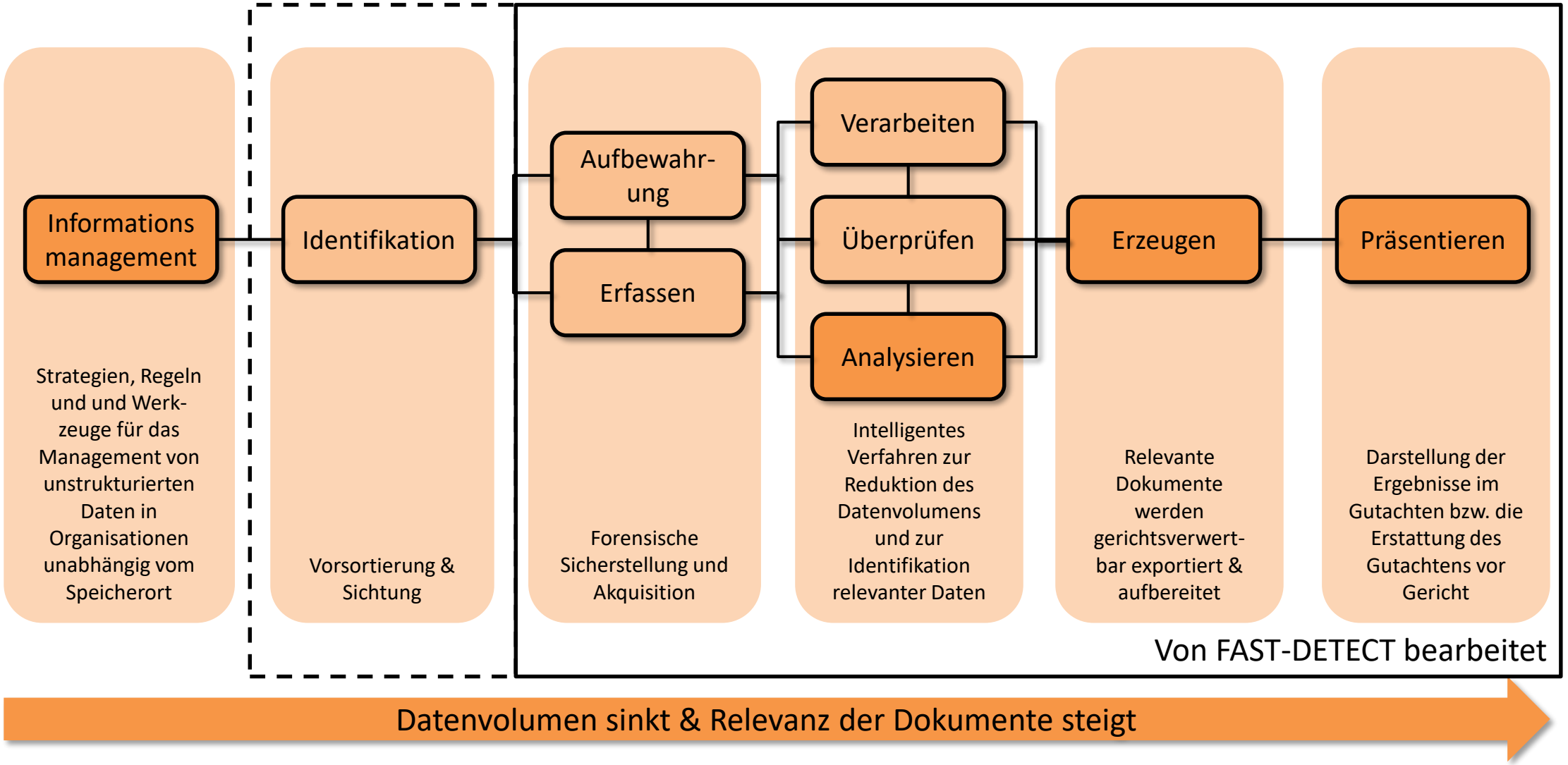
Viel wichtiger ist jedoch die korrekte Durchführung des komplexen eDiscovery Prozesses.

Clearwell

Relativity

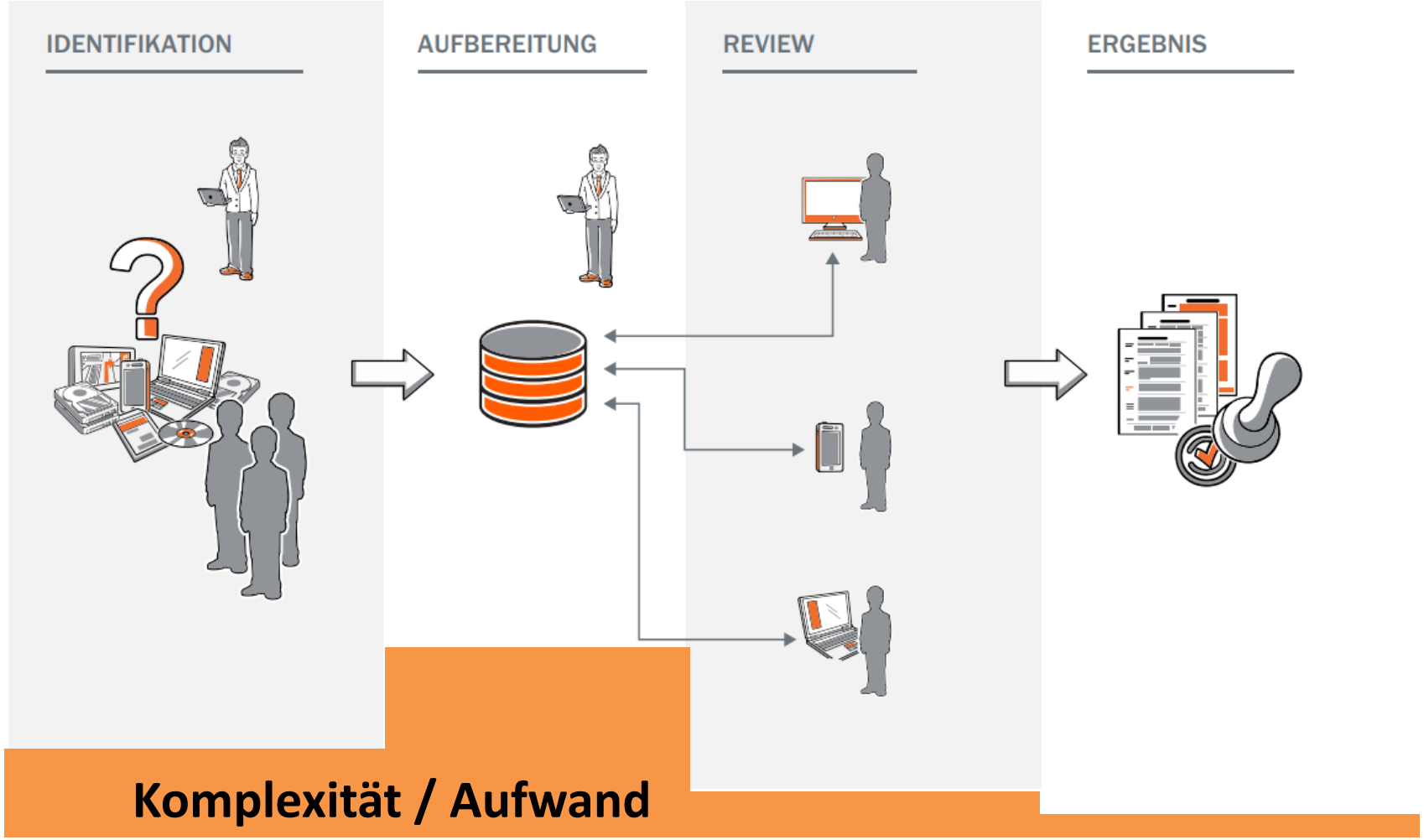
IBM Watson™

Das Electronic Discovery Reference Model



Datenvolumen sinkt & Relevanz der Dokumente steigt

Der FAST-DETECT eDiscovery Prozess



Identifikation

Informationsquellen		Resultat
Business Processes	• Anhand von Arbeitsprozessen kann festgestellt werden, mit welchen Systemen sachverhaltsrelevante Informationen verarbeitet wurden.	• Identifikation von relevanten IT-Systemen • Priorisierung von relevanten IT-Systemen bei der Datensicherung • Dokumentation der IT-Infrastruktur
Ablage-system	• Wie arbeiten Mitarbeiter im Unternehmen? Wo werden Daten, basierend auf den Arbeitsanweisungen, abgelegt? • Modus Operandi bei der Datenablage durch einen Mitarbeiter	
Informa-tionsma-nagement	• Vorhaltefristen? • Archivierungsfristen? • Dokumentenmanagementsysteme?	

Datensicherung

Strukturierte Daten

- Datensicherung mit dem Zweck des späteren Betriebs der Anwendung für die Analyse
- i.d.R. ein Take-All Prinzip
- Systemabhängige Datensicherung

Unstrukturierte Daten

- Selektive Sicherung relevanter Daten
- Datensicherung aus verschiedensten Datenquellen
  - Archivsysteme
  - Daten-Backups
  - E-Mail Systemen

Forensische Datensicherung zur Sicherung der Datenintegrität und Nachvollziehbarkeit

Typische Probleme bei der Datensicherung

- Extraktion der relevanten Daten aus speziellen (Dokumentenmanagement-)Systemen
- Forensische Sicherung von Cloud- und Webbasierten E-Mail-Angeboten
- Umgang mit Verschlüsselung auf Client-Computern
- Fehlende Schnittstellen zum Anschluss von Datensicherungshardware

Mit einer unvollständigen Datensicherung ist das Projekt vielleicht schon beim Start gescheitert.

Verarbeiten

Identifizierte unstrukturierte Daten

(eDiscovery Software)

1) Indizierung: Erkennung des Texts in den Dokumenten (Extracted Text)
2) Extrahierung der Metadaten der Dokumente
3) Erstellung von Dokumentenfamilien / Hierarchien

(eDiscovery Software)

Normalisierte Form aller unstrukturierten Daten - die Daten sind durchsuchbar und unabhängig von der Datenquelle darstellbar

Typische Probleme bei der Verarbeitung

- Die Verarbeitung von verschlüsselten Dokumenten ist nicht möglich.
- Exotische Datenformate können nicht interpretiert werden.
- Die Zulieferung von externen Datenquellen (sogenannter „Loadfiles“) ist fehlerhaft.
- Technische Programmefehler führen zu invaliden Ergebnissen / Datenbankfehlern.

Eine u.U. mehrtägige oder mehrwöchige Datenverarbeitung wird ungültig und muss wiederholt werden.

Überprüfen

Nach der Verarbeitung müssen folgende Sonderfälle überprüft werden:

- Sind Fehler bei der Verarbeitung aufgetreten?
- Wurden *alle* Daten verarbeitet?
- ***Plausibilitätscheck der Daten***
- Existieren verschlüsselte Dokumente?
- Müssen gewisse Dateitypen (PDF, TIFF etc.) einer Texterkennung unterzogen werden?

Arten von Suchen

Stichwortsuchen

Bsp.: Kickback

Stichwortsuchen mit booleschen Operatoren

Bsp.: Kickback AND („Igor“ OR „Poppov“ OR „IPO“)

Musterabgleich (Pattern Matching)

Suche nach allen Zahlenfolgen, die z.B. wie eine Kreditkartennummer aussehen.

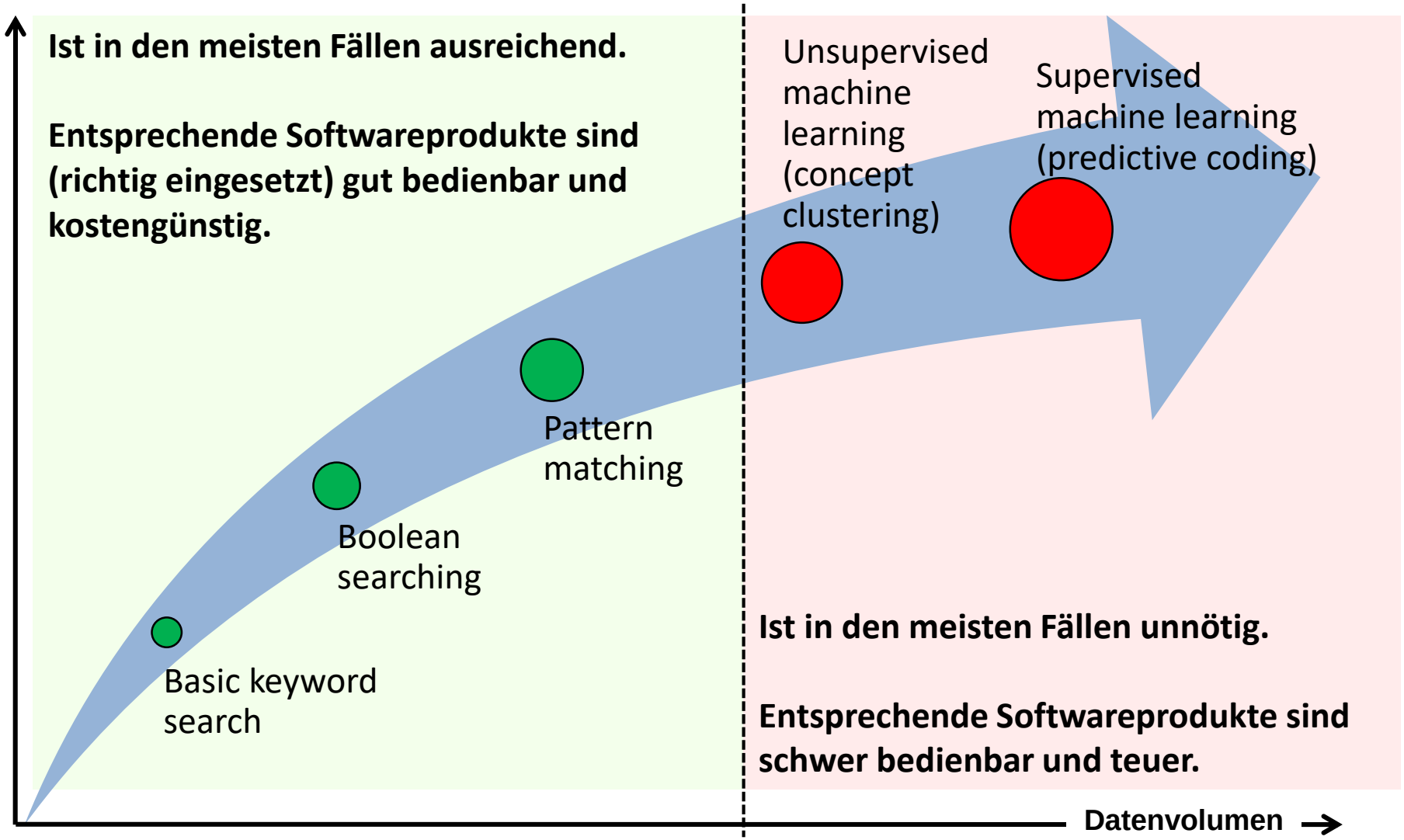
Clustering

Automatische Gruppierung von Suchergebnissen in thematische Gruppen

Predictive Coding

Unterstützung von maschinellem Lernen zur automatischen Einordnung von Suchergebnissen bezüglich Relevanz

Arten von Suchen



Strategie: Suchwortlisten

1. Definition von Suchworten oder Suchen

2. Test: Review einer kleinen Menge an Dokumenten und Prüfung der Treffer
   Fortlaufende Qualitätssicherung der Suchwortliste

3. Statistische Überprüfung, ob in den Dokumenten, die nicht auf die in (1) definierten Suchen anspringen, relevante Dokumente enthalten sind.

4. Review der Dokumente, die auf die Suchwortliste anspringen.
   Es muss kontinuierlich überprüft werden, ob die in (1) definierten Suchwörter abhängig von den Ergebnissen angepasst werden müssen.

Wahl der Stichprobe

Population	Fehlerspanne			Konfidenzniveau		
	10 %	5 %	1 %	90 %	95 %	99 %
100	50	80	99	74	80	88
500	81	218	476	176	218	286
1.000	88	278	906	215	278	400
10.000	96	370	4.900	264	370	623
100.000	96	383	8.763	270	383	660
> 1.000.000	97	384	9.513	271	384	664

Tagging / Coding Layout

Während einer Review werden die Dokumente abhängig von ihrer Relevanz und weiteren Kriterien kategorisiert. Dies geschieht über sogenannte Tags. Einem Dokument können während der Review beliebig viele Tags zugeordnet werden.

Abhängig von der Review und dem Ziel sollte das Coding Layout erstellt und auch während der Review möglicherweise angepasst werden.

Einfaches Coding Layout

1. Relevant
2. Nicht relevant

Komplexeres Coding Layout

Pflicht (einfach Auswahl)	Optional (mehrfache Auswahl)
Relevant	Sachverhalt 1
Nicht relevant	Sachverhalt 2
Korrupt	Sachverhalt 3
Weiterleitung	
Priviledged / Confidential	

Durchsicht (Review)

Return to document list

EN008648

Document 3 of 16

Viewer Native Extracted Text

Image

Edit Coding

100%

Badeer, Robert

From: carla.hoffman@enron.com

Sent: Tue, 8 Aug 2000 04:39:00

To: tim.belden@enron.com, robert.badeer@enron.com, jeff.richter@enron.com, phillip.platter@enro
mike.swerzbin@enron.com, diana.scholtes@enron.com, sean.crandall@enron.com,
matt.motley@enron.com, mark.guzman@enron.com, tom.alonso@enron.com,
mark.fischer@enron.com, stewart.rosman@enron.com, greg.wolfe@enron.com,
kristian.lande@enron.com, monica.lande@enron.com, valarie.sabo@enron.com

Cc:

Bcc:

Subject: DJ Power Price Cap In Calif Puts Brinksmanship On Grid

---------- Forwarded by Carla Hoffman/PDX/ECT on 08/08/2000 11:45
AM ----------

Enron Capital & Trade Resources Corp.

From: "Pergher, Gunther" <Gunther.Pergher@dowjones.com>
08/08/2000 11:34 AM

Attorney

66.07 kb/s

Reviewer Details

Control Number:	EN008648
Custodian:	Badeer, Robert

Coding

Responsiveness:	Responsive
Confidentiality:	Not Confidential
Privilege:	Not Privileged;
Issues:	Hot;

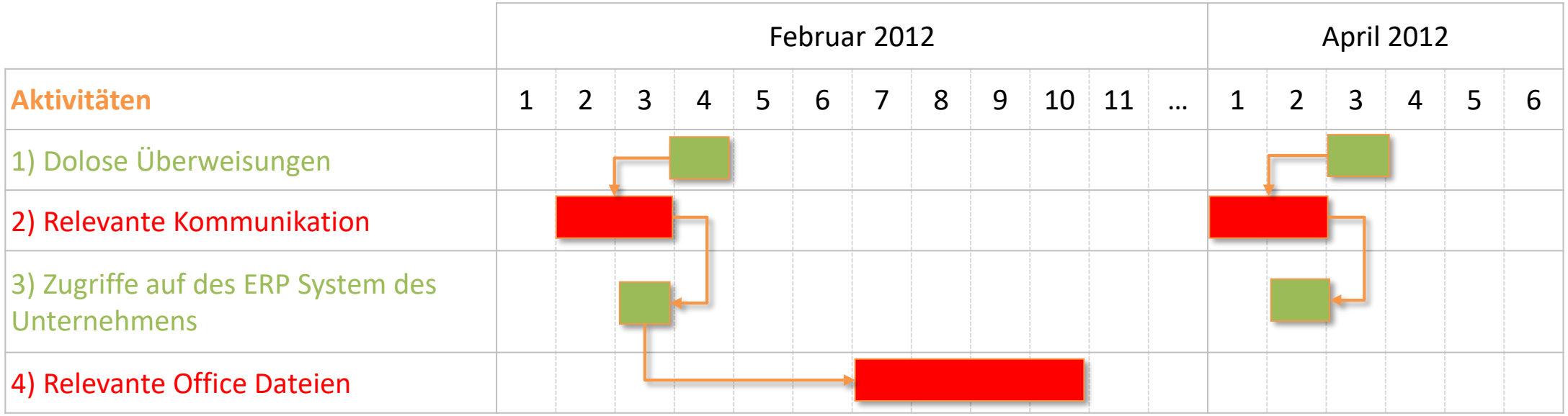
Reviewer Comments

Edit

Family

No Family

Verknüpfung von strukturierten & unstrukturierten Daten



Legende

- Strukturierte Daten
- Unstrukturierte Daten

In der IT spricht man von ***strukturierten Daten***, wenn die Daten in einer Datenbank als Felder mit bestimmtem Inhalt innerhalb einer Datenbankstruktur gespeichert sind (z.B. Felder für ein Überweisungsdatum und einen Betrag in der Buchhaltung).

Unstrukturierte Daten sind hingegen Briefe, Vertragsdokumente oder E-Mails bei den sich der Zusammenhang zu bestimmten Vorgängen nur aus dem textlichen Inhalt ergibt.
Während klassische e-Discovery-Methoden es nicht ermöglichen beide Informationsarten miteinander in Beziehung zu bringen, ist es mit den Methoden der IT-Forensik möglich z.B. über Zeitstrahlanalysen Beziehungen zwischen strukturieren und unstrukturierten Daten aufzuzeigen.

Fallbeispiel: Dokumentenrecherche

Auftrag

Der Beschuldigte steht unter Verdacht verschiedene Dokumente gefälscht zu haben. Diese Dokumente dienten zur Entlastung des Beschuldigten in einem Betrugsverfahren. Zur Auswertung standen sowohl der private als auch der geschäftlich genutzte Rechner zur Verfügung.

Vorgehen

- Identifikation von versteckten Dateien
- Einbinden aller E-Mails und Anhänge in die Suche
- Hashwert-Berechnung zum Abgleich identischer Kopien
- Volltextsuche nach relevanten Begriffen aus den fraglichen Dokumenten inkl. Suche nach Kontonummern (Pattern Matching)

Ergebnisse

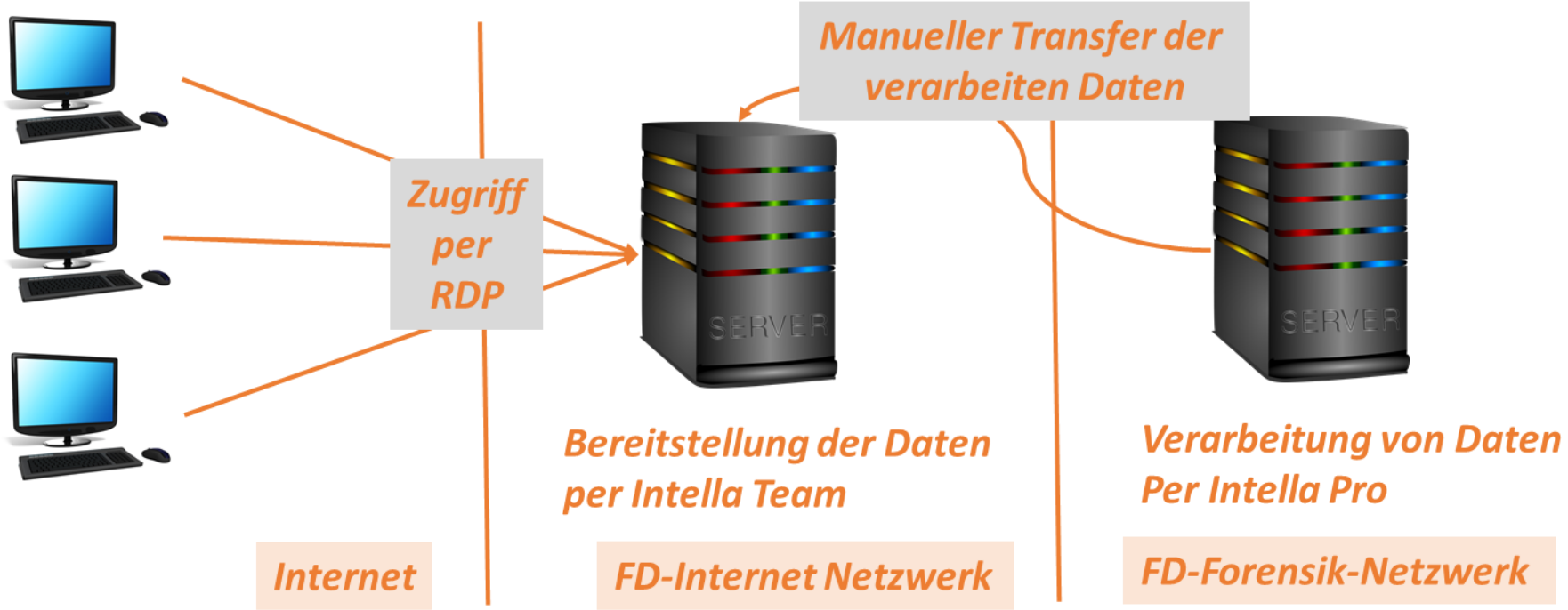
- E-Mail an die Geschäftsadresse des Beschuldigten auf dem Privatrechner im Ordner „Gesendet“ enthält das Dokument „Doc1.doc“ als Anhang: Nach einigen unauffälligen Seiten finden sich alle fraglichen Schreiben - jeweils ohne Briefkopf
- Inhaltlich gleiches Dokument (mit abweichendem Druckdatum) befindet sich auch im Internetcache des Arbeitsrechners

Befund

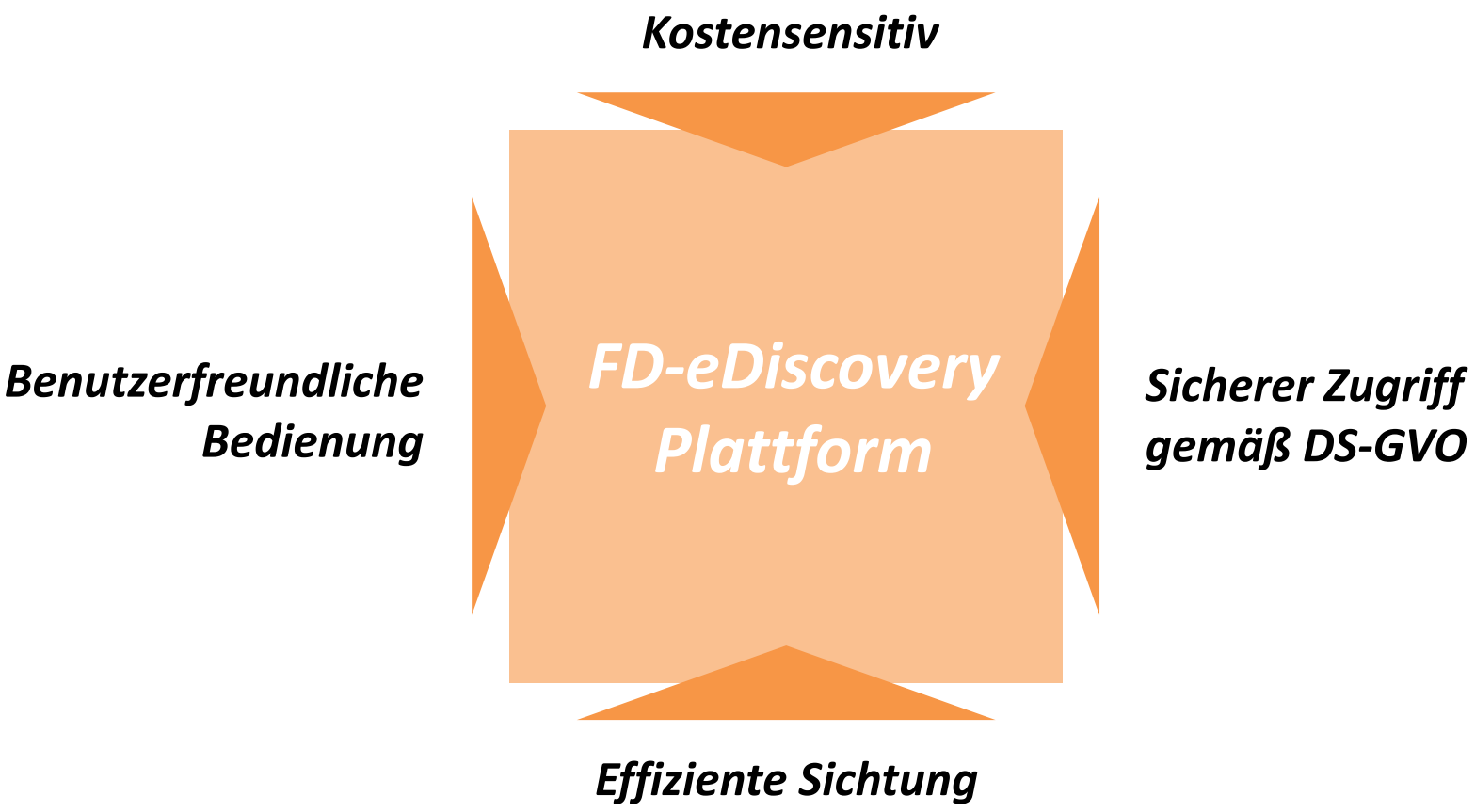
Über die Dokument-Metadaten ließ sich belegen, dass das Dokument am frühen Morgen zu Hause bearbeitet, dann per E-Mail versendet und vom Arbeitsplatz aus (nach Aufruf in einer Webmail-Oberfläche) gedruckt wurde.

Technische Details zur eDiscovery-Infrastruktur

Hierbei handelt es sich nicht um eine Web-Review-Plattform, sondern um eine Plattform, mit der Sie per Fernzugriff arbeiten können.



Die Ziele der FD-eDiscovery Plattform



Referenzen im Bereich eDiscovery (2018 / 2019)

- 15 eDiscovery Projekte verschiedener Größen in 2018 & 2019
- Kunden aus der Justiz, Rechtsanwaltskanzleien & Unternehmen
- Gesamtprozess inkl. Sicherung vor Ort, Aufbereitung und Bereitstellung
- Typisches Datenvolumen: ca. 10 – 200 GB

- Beispiel: Großprojekt
  - IT-forensische Datensicherungen für 600 Einzelgeräte
  - Gesichertes Datenvolumen ca. 180 TB
  - In der Reviewumgebung bereitgestelltes Volumen ca. 30 TB
  - Über 100 Millionen Einzelitems (1 Item = E-Mail / Dokument / Bild)

Ansprechpartner z.B. bei Kanzleien nennen wir gerne auf Anfrage

Live – Demonstration

Vielen Dank für Ihre Aufmerksamkeit

FAST-DETECT GmbH

Inselkammerstr. 12
82008 Unterhaching

Tel. +49 89 204040-0
Fax +49 89 204040-299
info@fast-detect.de
www.fast-detect.de

Geschäftsführer

Thomas Salzberger
Thomas.Salzberger@fast-detect.de

Dominic Degel
Dominic.Degel@fast-detect.de