Follow Along

Die Folien finden Sie unter:

https://kolloquium.teismar.de

Bachelorarbeit Kolloquium

Detection and Classification of Phishing Sites by Analyzing Common Patterns

Leibniz-Fachhochschule Hannover
In Kooperation mit Deutsche Telekom Security GmbH

Tim Julian Eismar | Matrikelnummer: 45768

21. Juli 2025

Let's go phishing →

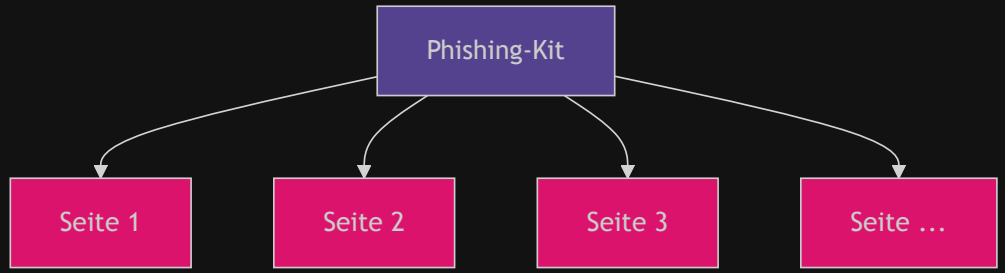
Gliederung

1. Motivation
2. Methodischer Überblick
3. Praktische Umsetzung
4. Ergebnisse
5. Fazit und Ausblick

Das Phishing-Problem

Aktuelle Bedrohungslage

- Zunehmende Anzahl und Raffinesse von Phishing-Angriffen
- Phishing-Kits ermöglichen einfache Erstellung betrügerischer Websites
- Schwierigkeit bei der Zuordnung neuer Seiten zu bekannten Kits

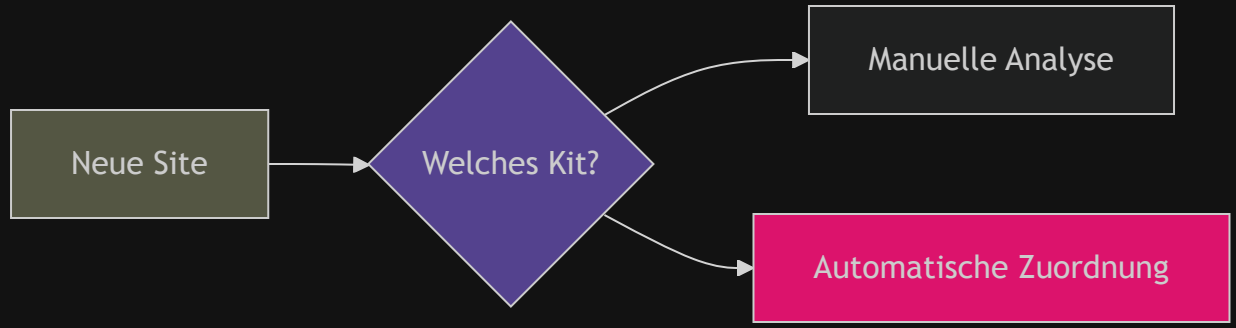


Zielsetzung

Was wollen wir erreichen?

- **Automatische Klassifizierung** neuer Phishing-Sites
- **Actionable Intelligence**
- **Schnelle Reaktion** auf neue Bedrohungen

Die zentrale Herausforderung



Forschungsfragen

1. **Merkmalsdefinition**
   Welche Merkmale können zur Identifizierung von Phishing-Kits verwendet werden?

2. **Merkmalsextraktion**
   Wie können diese Merkmale automatisch extrahiert werden?

3. **Matching-Verfahren**
   Wie können extrahierte Merkmale zur Zuordnung verwendet werden?

Methodischer Überblick

1. Merkmalsdefinition

- Analyse bestehender Phishing-Kits
- Identifikation gemeinsamer Charakteristika

2. Merkmalsextraktion

- Automatisiertes Aufsetzen von Phishing-Websites
- Feature-Extraktion mittels Web-Scraping

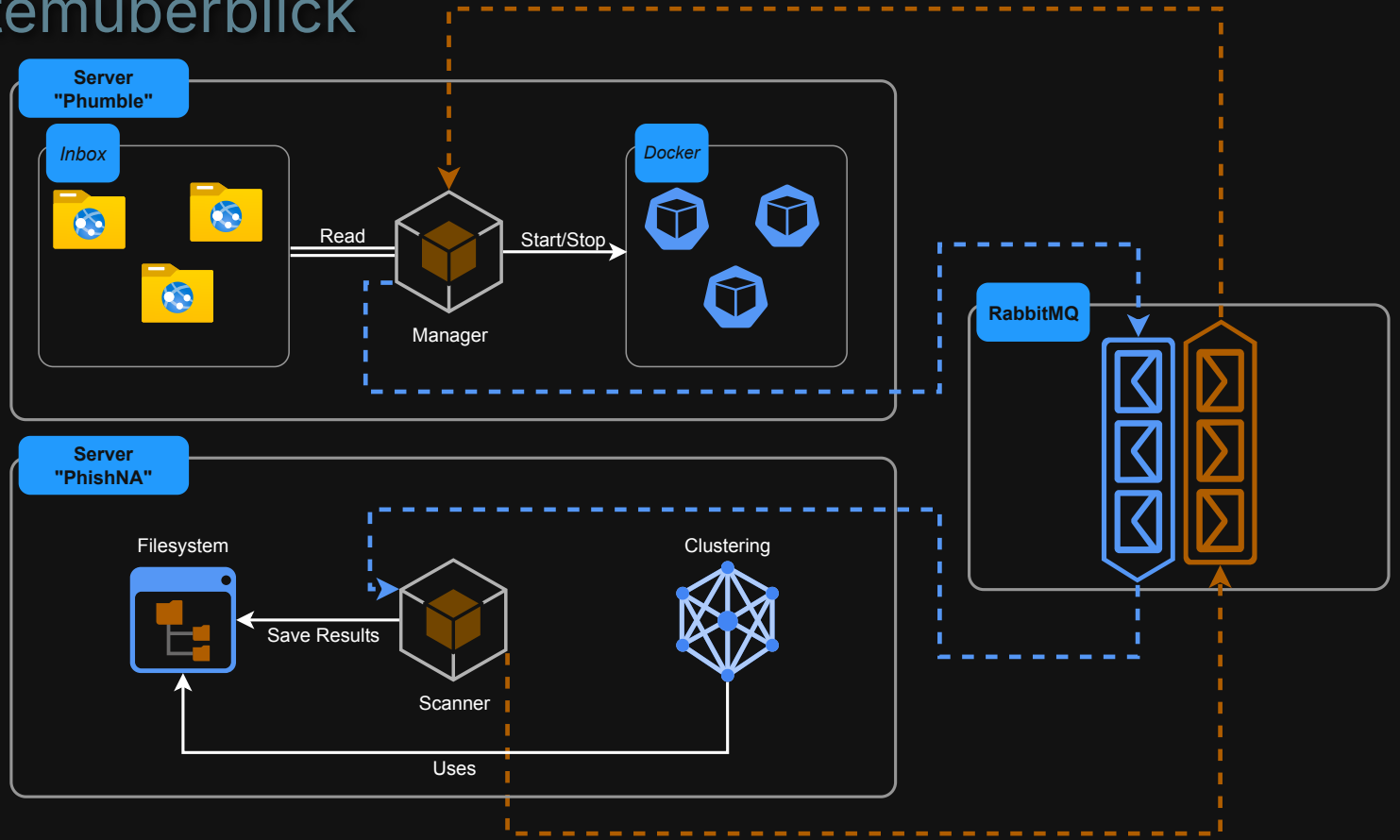
3. Matching

- Preprocessing der Merkmale
- Anwendung von Clustering-Algorithmen (DBSCAN)

4. Evaluation

- Manuelle Validierung der Ergebnisse

Systemüberblick



DBSCAN - Warum diese Wahl?

Vorteile für Website-Analyse

- Beliebige Cluster-Formen (nicht nur kreisförmig)
- Robust gegen Rauschen und Ausreißer
- Automatische Cluster-Anzahl
- Keine Vorab-Annahmen über Cluster-Form

Besonders wichtig für Phishing-Sites:

- Erkennung unterschiedlicher Kit-Strukturen
- Automatische Outlier-Erkennung
- Keine Vorab-Cluster-Anzahl erforderlich

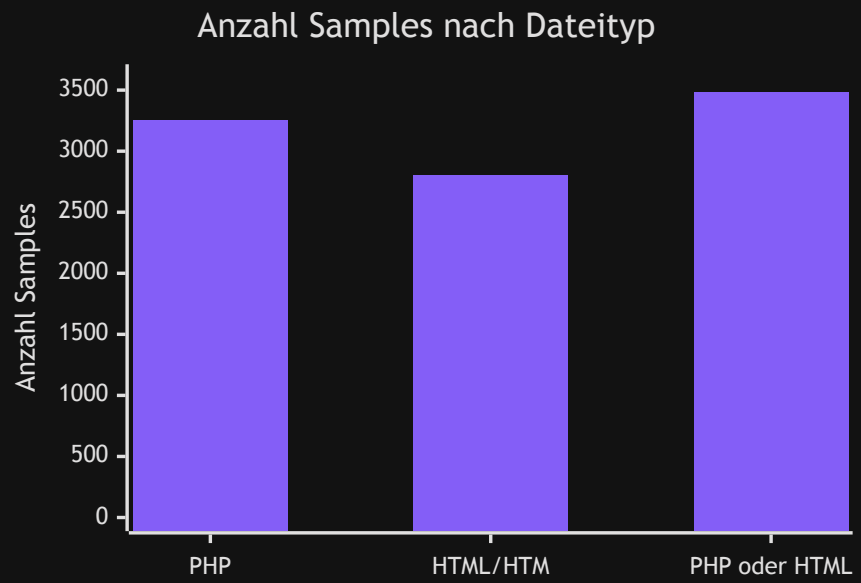
Datensammlung - Quellen

Woher kommen die Daten?

- **VirusTotal** (Premium API)
- **UrlScan** (Premium Zugang)
- Phishing-Repositories
- Directory Listings

Fokus auf PHP/HTML

- **3.900** Phishing-Samples analysiert
- **3.481** enthalten PHP oder HTML
- **462** Samples für Experimente verwendet



Feature-Extraktion - HTML

Strukturelle Merkmale

- **Link-Anzahl** (`link_count`)
- **Formular-Felder** (`form_count`)
- **Script-Tags** (`script_count`)
- **Passwort-Eingaben** (`password_input_count`)

Inhaltliche Merkmale

- **Sichtbare Textlänge** (`visible_text_length`)
- **Keyword-Häufigkeiten** (login, password, verify)
- Meta-Tags und Kommentare

Feature-Extraktion - HTTP

Response-Charakteristika

- Status-Codes
- Response-Größe
- Antwortzeit
- Content-Type Headers

Verhalten

- Cookies und Sessions
- Redirect-Verhalten
- Error-Handling

Feature-Extraktion - Scoring

Bewertungsmethoden

Kategoriale Entropie:
$S = -\sum_i p_i \log_2(p_i)$

Numerische Varianz:
Streuung der Merkmalswerte

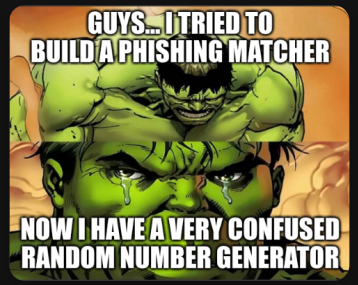
Gewichtete Gesamtbewertung:
Kombination beider Maße

Datenvorverarbeitung

- One-Hot-Encoding für kategoriale Features
- Standardisierung numerischer Features
- PCA für Dimensionsreduktion

Let the machine learn

- 50 Pfade `/index.php`, `/login.php`, etc.
- Lange wartezeit
- uuuuund...



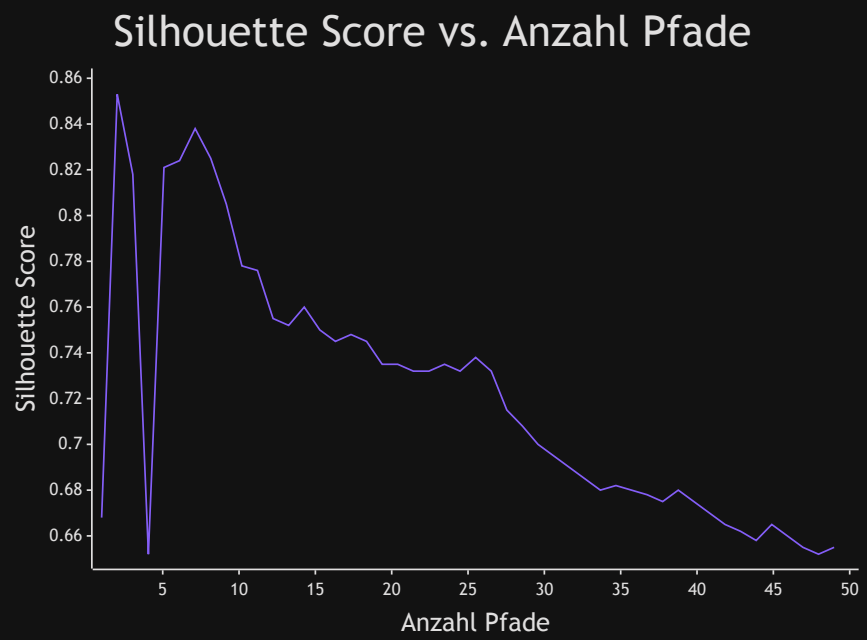
Experimentelle Ergebnisse - Pfadanzahl

Iterative Analyse

- **Evaluation für 1-50 Pfade**
- Silhouette-Score als Qualitätsmaß

Beobachtungen

- **1 Pfad:** Unzureichende Information
- **2 Pfade:** Optimale Balance
- **>2 Pfade:** Curse of Dimensionality



Clustering-Ergebnisse (2 Pfade)

Cluster-Verteilung

Cluster	Anzahl Samples	Charakteristikum
-1	4	Rauschen
0	156	Fehlende Index-Dateien
1	259	Malformierte Sites
2-6	4-19	**Logische Gruppierung**

Beobachtung

Kleinere Cluster (2-6) zeigen **hohe Kohärenz** und sinnvolle Gruppierung

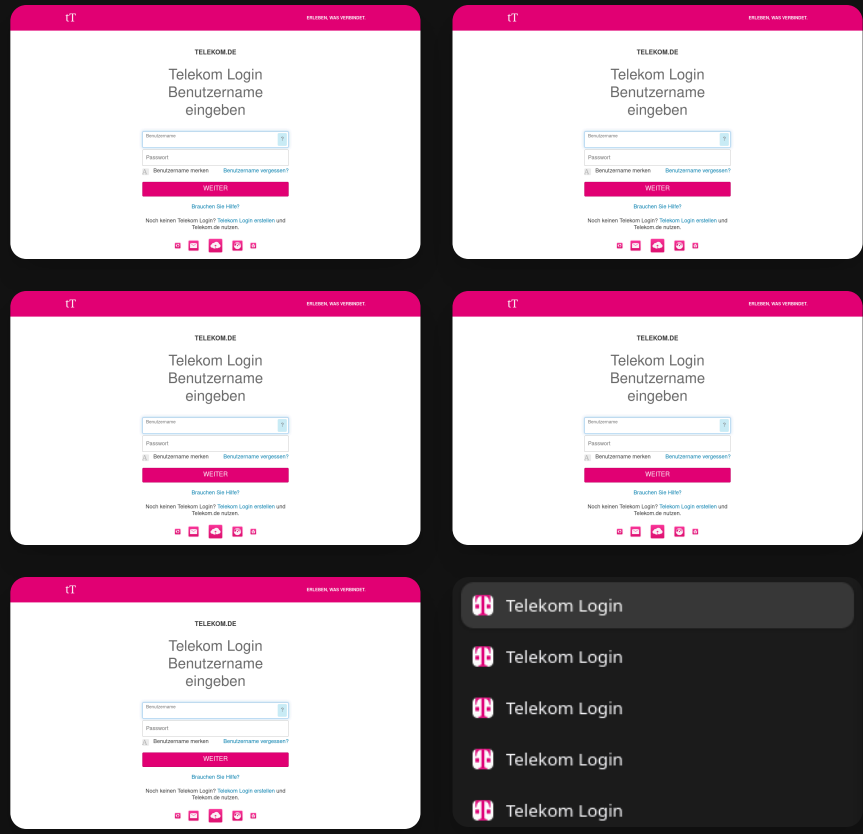
Erfolgreiche Cluster-Analyse

Cluster 6: Telekom Login

- **5 identische Phishing-Sites**
- Alle imitieren Telekom Login-Seite
- **Visuell identische Darstellung**
- Hohe Cluster-Kohärenz

Validierung bestätigt:

- Gemeinsame HTML-Struktur
- Identische Styling-Elemente
- Gleiche Formular-Felder



Analyse der großen Cluster

Cluster 0: Strukturelle Anomalien

- Websites **ohne Standard-Index-Dateien**
- Fehlende `index.html` oder `index.php`
- Content in Subdirectories
- **Evasion-Strategie**

Cluster 1: Anti-Bot Mechanismen

- Malformierte oder fehlerhafte Sites
- CAPTCHA und Bot-Protection
- **Defensive Mechanisms**

Bewertung - Erfolge

Was funktioniert gut:

- ✅ **Automatische Feature-Extraktion** funktioniert
- ✅ **DBSCAN** identifiziert sinnvolle Cluster
- ✅ **Hohe Kohärenz** in kleineren Clustern
- ✅ **Strukturelle Anomalien** erkannt
- ✅ **Evasion-Techniken** identifiziert

Das System identifiziert sowohl strukturelle Ähnlichkeiten als auch Evasion-Techniken

Bewertung - Herausforderungen

Verbesserungsbedarf:

- ⚠️ **Große Cluster** schwer interpretierbar
- ⚠️ **Limitierte Datenmenge** (462 Samples)
- ⚠️ **PCA** erschwert Feature-Interpretation
- ⚠️ **Manuelle Validierung** erforderlich

Aber dennoch:

Grundlegende Machbarkeit **erfolgreich demonstriert**

Antworten auf Forschungsfragen

1. Identifikation von Merkmalen

- **HTML-Struktur** (Links, Forms, Scripts)
- **HTTP-Verhalten** (Status, Response-Size)
- **Dateihierarchie** und Pfad-Struktur

2. Automatische Extraktion

- Docker-basiertes Aufsetzen von Phishing-Sites
- Web-Scraping Pipeline
- Strukturierte Feature-Extraktion

3. Matching-Verfahren

- **DBSCAN-basierte Clustering**
- **PCA** für Dimensionsreduktion
- **Silhouette-Score** Optimierung

Fazit - Hauptergebnisse

Was haben wir erreicht?

- ✅ **Funktionsfähiges System** zur Phishing-Kit-Analyse
- ✅ **Strukturelle Muster** erfolgreich identifiziert
- ✅ **Automatisierte Pipeline** entwickelt
- ✅ **Validierung** durch manuelle Inspektion

Beitrag zur Cybersicherheit

Grundlage für **automatisierte Threat Intelligence**

Ausblick - Methodische Verbesserungen

Nächste Schritte

- **Erweiterte Feature-Sets**
- **Verhaltensbasierte Merkmale**
- **Supervised Learning Integration**
- **Realzeit-Erkennung**

Technische Erweiterungen

- **Größere Datensätze** (Tausende von Samples)
- **Verbesserte Visualisierung**
- **Performance-Optimierung**

Ausblick - Praktische Anwendung

Deployment Möglichkeiten

- **Integration in SOC-Systeme**
- **Kooperation mit Strafverfolgung**
- **Continuous Monitoring**
- **Threat Intelligence Feeds**

Forschungsrichtungen

- **Evasion-Techniken**
- **Multi-Language Support**
- **Cross-Platform Analysis**

Vielen Dank für Ihre Aufmerksamkeit!

Tim Julian Eismar

tim.eismar@telekom.de

Leibniz-Fachhochschule Hannover
Deutsche Telekom Security GmbH

Quellen

Dokumente & Recherche

- Meiner Bachelorarbeit zu entnehmen

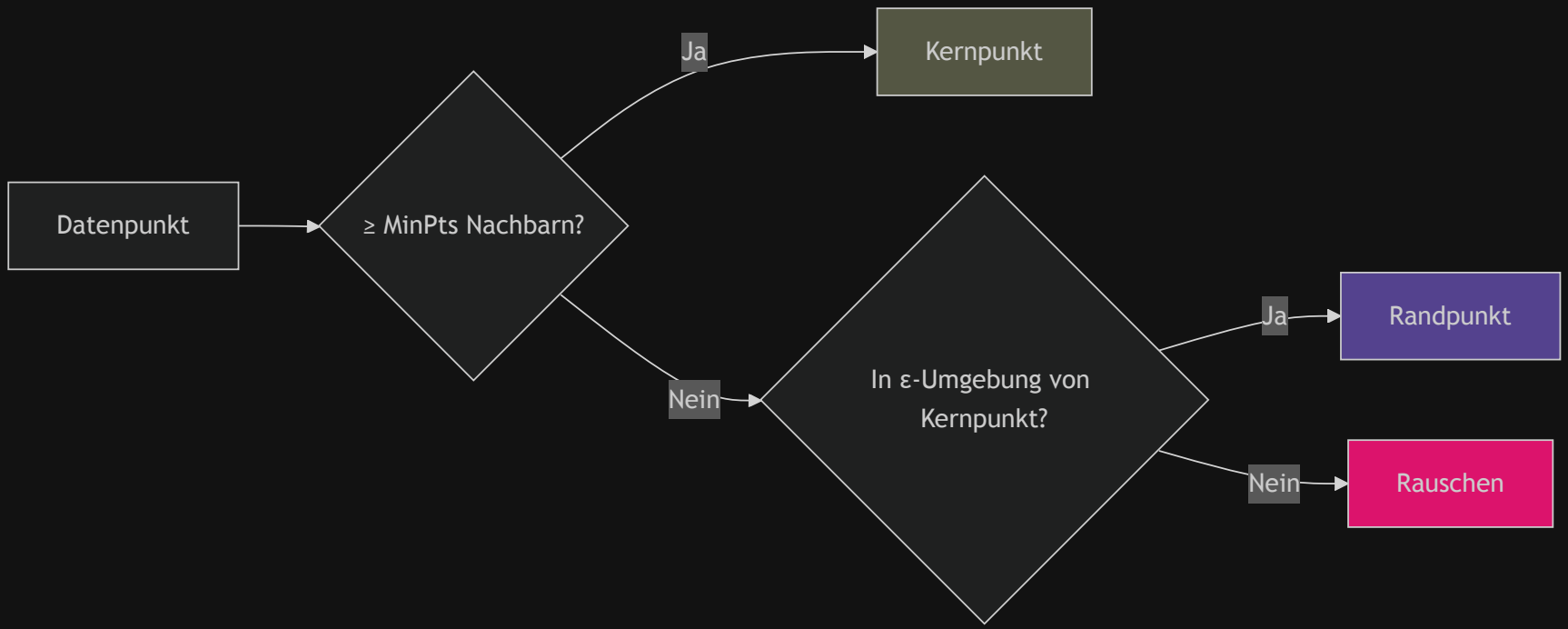
Bilder & Medien

- Unsplash
- Imgflip Meme Generator
- Google Imagen 4
- Medium
- Smartsoc Solutions

Fragen?


Backup Folien

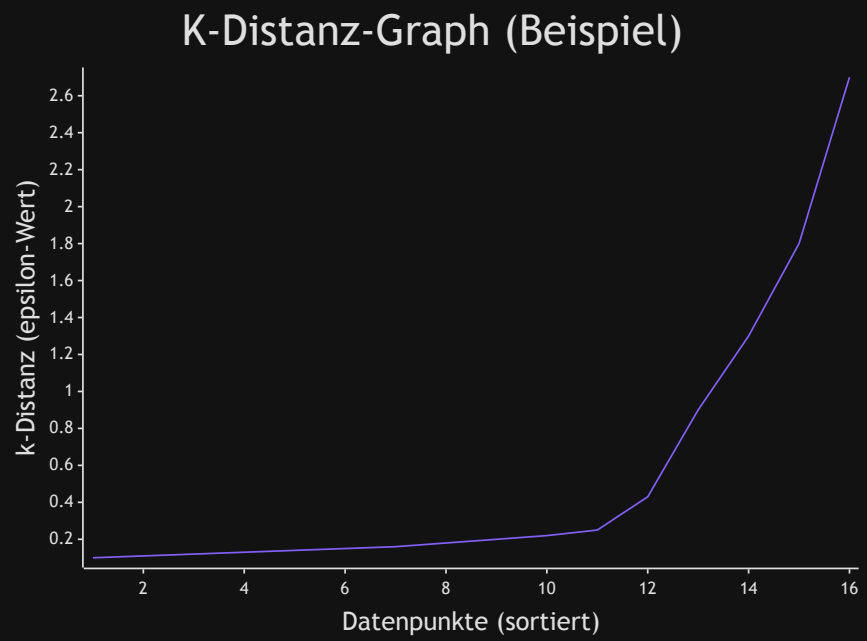
DBSCAN - Funktionsweise



Anhang: DBSCAN Parameterwahl (ε)

Bestimmung des Epsilon-Wertes (ε)

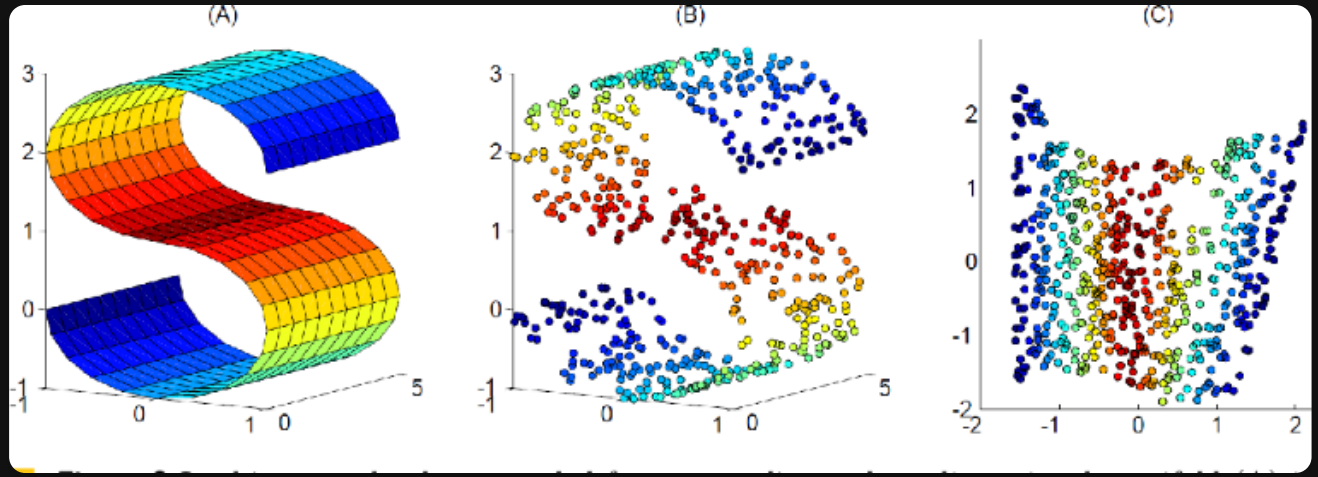
1. **Berechnung:** Für jeden Datenpunkt wird der Abstand zum k-nächsten Nachbarn berechnet (hier `k=MinPts`).
2. **Sortierung:** Die Abstände werden aufsteigend sortiert und aufgetragen.
3. **"Ellenbogen"-Kriterium:** Der optimale `ε`-Wert befindet sich am "Ellenbogen" (Knie) der Kurve. Dieser Punkt markiert den Übergang zwischen dichten Clustern (flacher Anstieg) und Rauschen (steiler Anstieg).



Curse of Dimensionality

Wenn Daten zu viele Merkmale (Dimensionen) haben, wird es schwierig:

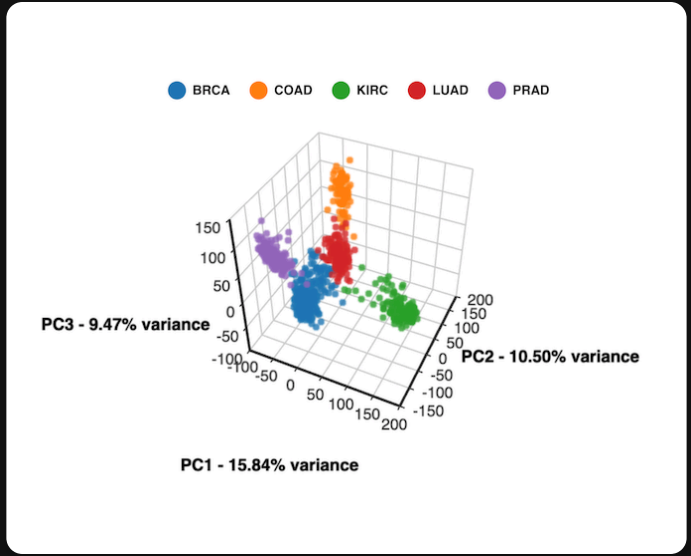
- **Algorithmen werden ineffizient:** Berechnungen dauern länger und werden unzuverlässiger.
- **Abstände verlieren an Aussagekraft:** Die Ähnlichkeit zwischen Datenpunkten wird unklar.
- **Visualisierung ist unmöglich:** Wir können Daten in mehr als drei Dimensionen nicht darstellen.



PCA & 3D-Visualisierung

Principal Component Analysis (PCA)

- **Vorgehen**: PCA transformiert die ursprünglichen, korrelierten Merkmale in einen neuen Satz von unkorrelierten Merkmalen, die Hauptkomponenten. Die ersten wenigen Komponenten enthalten den Großteil der Information.
- **Wie**: PCA berechnet die Eigenvektoren und Eigenwerte der Kovarianzmatrix der Merkmale und wählt die Komponenten mit den höchsten Eigenwerten aus.



Ethische Überlegungen

- **Umgang mit Daten:** Die analysierten Phishing-Kits können potenziell personenbezogene Daten (PII) von Opfern enthalten (z.B. in Log-Dateien). Die Analyse erfolgte auf von Außen unzugänglichen, anonymisierten Servern um die Exposition von PII zu vermeiden.
- **Keine aktive Interaktion:** Das System interagiert passiv mit den gehosteten Seiten und führt keine Aktionen aus, die Daten an die Phishing-Akteure senden könnten (z.B. Ausfüllen von Formularen).
- **Zweckbindung:** Die Analyse dient ausschließlich dem Zweck der Forschung und der Verbesserung von Verteidigungsmechanismen.

Methodische Limitationen

- **Datensatzgröße:** Die Analyse basiert auf 462 Kits. Obwohl die Ergebnisse vielversprechend sind, ist eine Validierung auf einem größeren Datensatz für eine Verallgemeinerung notwendig.
- **Fokus auf PHP/HTML:** Die aktuelle Methode ist auf PHP- und HTML-basierte Kits ausgerichtet. Moderne Phishing-Seiten, die stark auf JavaScript-Frameworks setzen, erfordern möglicherweise erweiterte Features.
- **Manuelle Validierung:** Die finale Bewertung der Cluster-Qualität erforderte eine aufwändige manuelle Inspektion der Ergebnisse.