Künstliche Intelligenz für sichere Web-Infrastrukturen mit digitalem Identitätsmanagement

Akronym: KIWI

(I) Kurzbericht zum Teilvorhaben "Föderiertes Lernen der Bewertung von Identitäten in föderierten Geschäftsprozessen"

Ansprechpartner:
Jan Griebsch
1&1 Mail & Media Development & Technology GmbH
Sapporobogen 6-8
E-Mail: jan.griebsch@1und1.de

1 Zielsetzung, Ausgangsbasis

Ausgangsbasis: Das Vertrauen der Nutzer in den Schutz ihrer digitalen Daten und Identitäten ist für 1&1 zentral. Dies steht jedoch im Spannungsfeld mit der komfortablen Nutzung (z.B. schwache Authentifizierung) und der Vermeidung von fälschlichen „Abuse"-Klassifizierungen, die das Vertrauen unterminieren können. Föderierte Geschäftsprozesse, monetär nutzbare digitale Identitäten und strenge Datenschutzregularien sind zusätzliche Anforderungen.

Zielsetzung: Entwicklung von Konzepten, Architekturen und KI-Verfahren, die die genannten Zielkonflikte durch die Nutzung transparent verfügbarer Metadaten (Protokolldaten, Nutzungsdaten usw.) verbessern. Ein föderierbarer Ansatz ist dabei wesentlich, bedingt durch Datenschutzgesetzgebung und die europäisch-mittelständischen Unternehmensstrukturen, die große, übergreifende Trainingsdatensätze unrealistisch machen.

2 Ablauf

Aufgrund der anhaltenden CoViD19-Pandemie und den damit verbundenen Einschränkungen verzögerte sich unser Projekt bei 1&1 erheblich. Es war anfänglich nicht möglich, unter den erschwerten Bedingungen rechtzeitig neue Mitarbeiter einzustellen, Arbeitsplätze einzurichten oder die benötigten Arbeitsmittel zu beschaffen. Diese Verzögerungen konnten bis zum geplanten Projektende nicht aufgeholt werden, weshalb eine kostenneutrale Projektverlängerung beantragt und genehmigt wurde.

In Workshops wurde mit den Partnern eine **Zielarchitektur** erarbeitet. Parallel dazu wurden darauf abzubildende **Use-Cases** definiert. Die bei den Projektpartnern sehr heterogen ausfallenden Teilvorhaben reichen von der Sicherung von Industrieanlagen, über die Analysen von DDOS oder OAUTH Netzwerkprotokollen bis zu der o.g. Sicherung digitaler Identitäten bei 1&1. Gespräche zu **Kooperationsmöglichkeiten** bei der DDOS Mitigation und zur OAUTH Protokollvulnerabilitätsuntersuchung blieben ohne konkretes Ergebnis.
KI-Verfahren hängen entscheidend von den zugrundliegenden Datenverteilungen ab. Um die Kooperationsmöglichkeiten zu verbessern, entschied sich 1&1 den Partnern einen **Datensatz aus dem 1&1 Schwerpunkt Use-Case 'Missbrauch von Identitäten'** zu Verfügung zu stellen. Diese Kooperationsinitiative war es **mit aufwendigen datenschutzrechtlichen Klärungen** verbunden, und die am Ende in aufwendigen, starken Anonymsierungsmaßnahmen resultierten – diese Maßnahme würde sich nicht regelmäßig wiederholen oder weiter skalieren lassen, zu dem war auch das Interesse verhalten. Daher hat sich 1&1 Mail & Media Development & Technology GmbH folgend fokussiert auf:

- Die algorithmische **Kooperation mit der Hochschule Karlsruhe (Prof. Oliver Waldhorst)** in der Domäne '**Fraud Detection & Prävention**'.
- Die Modellierung und die prototypische Implementierung eines KI-Systems zur Kommunikationsdaten bezogenen Vertrauensbewertung von Login-Versuchen.

Auch der **generelle Projektverlauf unterlag starken Datenschutzeinschränkungen und/oder -bedenken,** die nur teilweise gelöst werden konnten. Dieser Umstand schlug auch erheblich bei der prototypischen Entwicklung des Login-Bewertungssystems zu buche. Ein datenschutzrechtlich praktikabler Lösungsweg mit Klärung der Rahmenbedingungen, der Bewertung des Systemdesign und des Datenmanagements konnte mit hohen, so nicht erwarteten Aufwänden, jedoch gefunden werden.

3 Ergebnisse

- Entwurf eines **Konzepts einer föderativen Architektur gemeinsam mit den Partnern**, dazu für die Partner relevante Use-Cases definiert.
    - Juristische Prüfung und **Zurverfügungstellung anonymisierter Produkt- und Kundendaten an die Partner**.

- Entwicklung eines, auf **Differential-Privacy Methoden basierenden Ansatzes zum DSGVO-freundliche Synthese und Austauschen hochdimensionaler Metadaten.**

- Untersuchungen zur Domänen-übergreifenden (Usermanagement, Premium-Products) Abuse-Erkennung innerhalb der Mail&Media GmbH mit föderierten KI-Modellen ergaben ein deutliches Verbesserungspotential bei der Erkennung von Betrug e.g. beim Domain-Kauf.
    - "Verwendung von maschinellem Lernen zur Klassifizierung missbräuchlicher Registrierungen" (Bachelorarbeit)

- Untersuchungen zur Betrugserkennung mit **föderierten KI-Modellen in Kooperation mit der 1&1 Telecommunications SE**.
    - Konzept(e) zur Datenschutz-optimiertem Informationsaustausch in Kooperation mit der HKA.
    - Externes Gutachten zur DSGVO-Konformität.
    - **Potential Betrugsschadenverringerung um Millionen Euro / Jahr – Umsetzung in Verhandlung**.

- Prototypisches **KI-System zur Erkennung von Identitätsmissbrauch in der Erprobung bei GMX und web.de**.
    - Konzeptentwicklung unter Berücksichtigung von Aspekten zu Datenschutz, Risikomitigation, Monitoring, Qualitätskontrolle und Kundenfeedback.
    - Experimente und Evaluation verschiedener KI-Ansätze:
        - Account-lokale, Populations-globale, sowie hybride Modelle
        - "A Reconstruction-Based One-Class Approach to Login Classification" (Master-Arbeit)
    - Aufbau benötigter Infrastruktur, Softwareentwicklung und Einbindung in Produktivsystem (Login).
    - **Schutz von mehreren Tausend Accounts pro Woche vor missbräuchlichem Zugriff - unerkannt durch bisherige Systeme**.
    - Aktive Weiter-erforschung und –entwicklung des Ansatzes erfolgt.

Künstliche Intelligenz für sichere Web-Infrastrukturen mit digitalem Identitätsmanagement

Akronym: KIWI

(II) Eingehende Darstellung zum Teilvorhaben “Föderiertes Lernen der Bewertung von Identitäten in föderierten Geschäftsprozessen”

Ansprechpartner:
Jan Griebsch
1&1 Mail & Media Development & Technology GmbH
Sapporobogen 6-8
E-Mail: jan.griebsch@1und1.de

Inhaltsverzeichnis

1 Zielsetzung, Ausgangsbasis

Ausgangsbasis: Für 1&1 ist das Vertrauen ihrer Nutzer in den Schutz ihrer digitalen Daten und Identitäten Kernbestandteil der Geschäftsstrategie. Ein, im Internet inhärent anspruchsvolles Ziel, welches zusätzlich jedoch im Spannungsfeld mit anderen steht: Komfortable Nutzung zum Beispiel bedingt Kompromisse wie eine relativ schwache Authentifizierung („Industriestandard“: Passwort). Fälschliche „Abuse“-Klassifizierungen und entsprechende Account-Sperrungen andererseits, sind aus Nutzersicht quasi ein Serviceausfall, und unterminieren das Vertrauen in vergleichbarer Weise. Föderierte Geschäftsprozesse, geschäftlich und monetär unmittelbar nutzbare digitale Identitäten, sowie immer explizitere Datenschutzregularien, kommen

als Anforderungen hinzu.

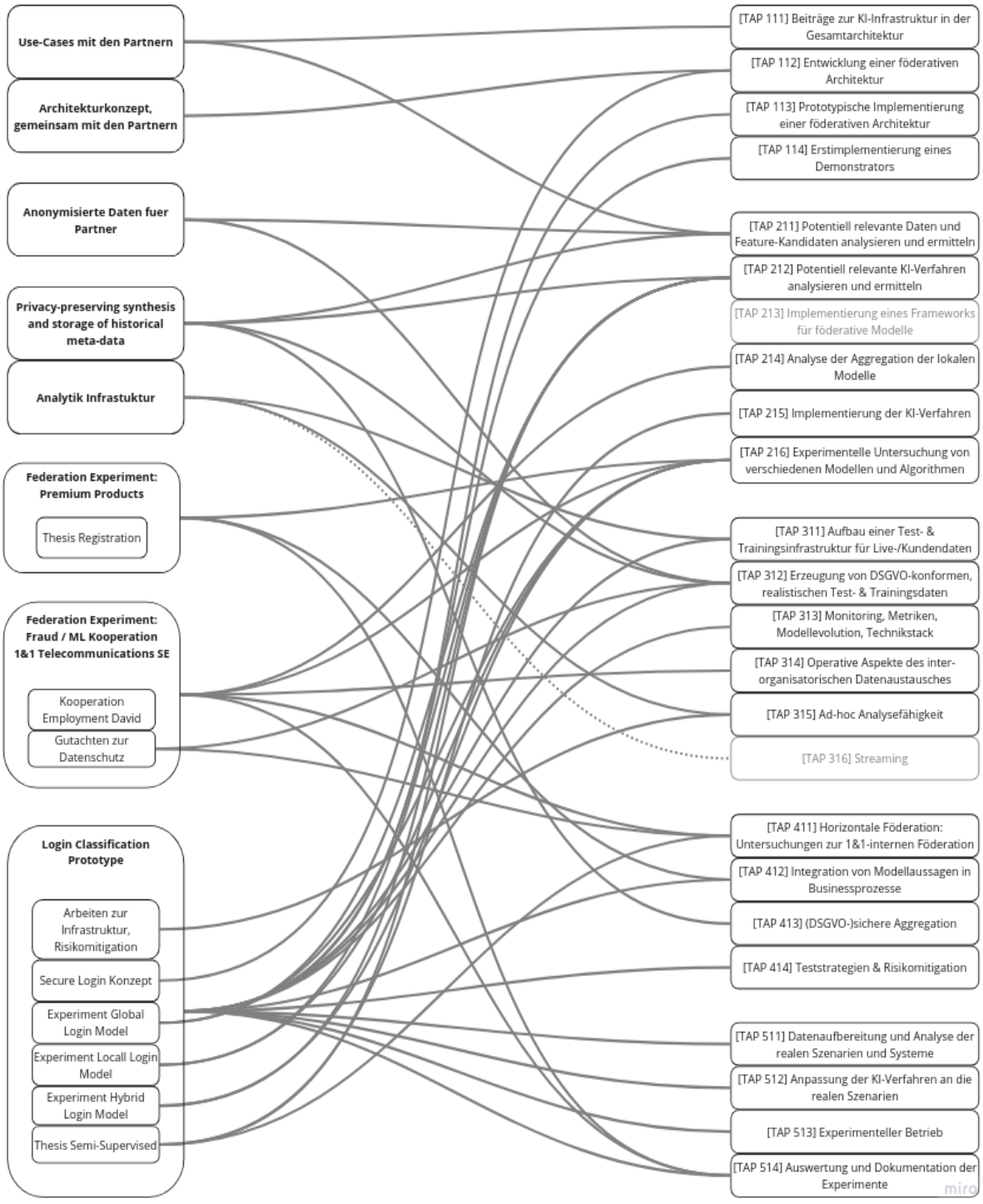
Zielsetzung: Entwicklung von Konzepten, Architekturen, und KI-Verfahren, die o.g. Zielkonflikte durch Nutzung transparent verfügbarer Metadaten (Protokolldaten, Nutzungsdaten usw.) verbessert. Grundsätzlich zu berücksichtigen sei dabei ein föderierbarer Ansatz. Bedingt ist diese zusätzliche Zielsetzung sowohl durch Datenschutzgesetzgebung als auch durch die typisch europäisch-mittelständischen Unternehmensstrukturen (die sich auch in der United Internet Holding und 1&1–intern widerspiegeln), die eine Zusammenführung in große, übergreifende Trainingsdatensätze unrealistisch erscheinen lässt.

1.1 Gesamtheitlicher Vergleich mit dem Vorhaben

Das von der Mail&Media GmbH durchgeführte KIWI-Projekt deckt sich thematisch weitgehend mit dem 2020 geplanten Vorhaben. Die ursprünglich vorgesehene Arbeitspaketstruktur erwies sich im Projektverlauf allerdings als nicht mehr passend und wurde dem realen Projektverlauf entsprechend angepasst.

Eine Hauptursache für diese Anpassung liegt in so, in diesem Umfang, im Vorhaben nicht berücksichtigten DSGVO-Aufwänden und –Einschränkungen. Dies ist umso bemerkenswerter als das (u.A.) DSGVO-Anforderungen dem Grunde nach als Motivation für einen der Forschungsschwerpunkte diente: der Föderation von KI-Systemen, die den Datenaustausch minimieren sollte. Eine breite, technisch-offene Föderation zum Schutz digitaler Identitäten erwies sich jedoch als – juristisch – unrealistisch. Das hat auch die Kooperationsmöglichkeiten mit den Vorhaben-Partnern eingeschränkt. Daher rückte im Projektverlauf fast zwangsläufig die Bearbeitung der Themen in Firmen-internen Kooperationen & Experimenten. Auch hier erwiesen sich DSGVO-Prüfungen und resultierende Bewertungsunklarheiten (DSGVO-Risiken) als ein Bremsklotz, der u.A. die Kooperation zweier, wirtschaftlich stark motivierter Tochterunternehmen um ca. 1Jahr verzögerte.

Dagegen waren die fachlich-technischen Ergebnisse der einzelnen KI-Anwendungsexperimente (Kapitel 6,7,8) für die Anwendungsfälle “Missbrauch von Identitäten “ sowie “Fraud Detection” sehr vielversprechend: Ein Projekt befindet sich in der Produkt-Umsetzung, über ein Weiteres wird auf Vorstandsebene verhandelt.



Vergleichende Darstellung der Bearbeitung der Vorhabenarbeitspakete (rechts) und -themen im realen Projektverlauf (links). Abhängigkeiten zwischen Paketen wurden zugunsten der Lesbarkeit nicht eingezeichnet.

2 Architekturkonzept und Anwendungsfälle mit den Partnern

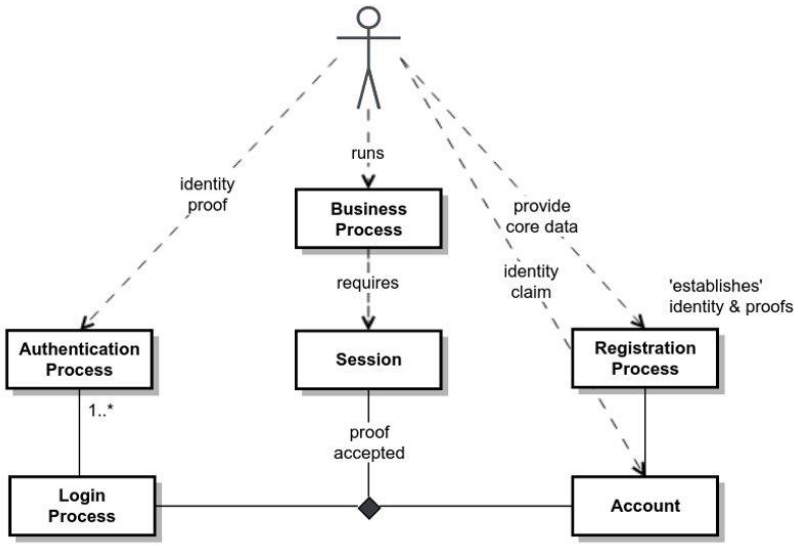
Dem Vorhabenplan folgend wurden mit den Partnern zunächst die Vorhaben Arbeitspakete (AP) 11, 112, sowie 211 bearbeitet. Das Vorgehen folgt klassischen Software-Projekt-Planungs-Prinzipien, nach denen Anwendungsfälle und dann eine passende Software-Architektur abgestimmt und definiert wird.

2.1 Anwendungsfälle

Mit den Partnern wurden die folgenden Anwendungsfälle definiert. Erkennbar ist eine ausgeprägte Heterogenität nicht nur den Anwendungs-Domänen, sondern auch bzgl. der fokussierten Angriffstypen und der Frage, inwieweit *Digitale Identitäten* primärer Untersuchungsgegenstand sind.

Anwendungsfall #1 Missbrauch von Identitäten (1&1):

Der Use Case fokussiert die Erkennung von Missbrauch digitaler Identitäten in Geschäftsprozessen. „Missbrauch“ ist die illegitime Annahme einer Identität (Impersonation) oder deren illegitime Verwendung. Föderierte Verfahren werden für das Training und die Inferenz von ML-Detektoren genutzt. „Föderierung“ bedeutet hier die Verteilung der Bewertung von Domänenentitäten auf verschiedene Organisationen. Ein Benutzer beansprucht eine Identität und wird für eine Session authentifiziert, die aus Anfragen mit Meta-Daten besteht. Es wird entschieden, ob die Identität und jeder Request legitim oder missbräuchlich ist, und der Zugriff auf Prozesse und Ressourcen wird entsprechend eingeschränkt. Dabei wird der Ablauf einer Session im authentifizierten Kontext betrachtet, nicht jedoch die Absicherung von Zugangsdaten oder Angriffe auf den Authentifizierungsprozess.



Domänenmodell für Anwendungsfall 1.

Anwendungsfall #2 Netzangriffe (KIT):

Dieser Use Case erforscht den Einsatz föderierter maschineller Lernverfahren zur Erkennung und Abwehr von Angriffen auf die Netzinfrastruktur von Web-basierten Identitätsmanagementsystemen. Betrachtet werden (Distributed) Denial-of-Service-Angriffe und typische Informationsbeschaffungsangriffe wie Port-Scans. Der Fokus liegt auf föderiertem Training von ML-Detektoren zur Abwehr von Netzangriffen basierend auf heterogenen Daten verschiedener Organisationen.

Anwendungsfall #3 Fraud Detection (adesso):

Dieser Use Case befasst sich mit der Erkennung von Kreditkartenbetrug, bei dem gestohlene Kreditkartendaten für unrechtmäßige Transaktionen genutzt werden. Ein föderierter Ansatz wird angestrebt, bei dem Zahlungsdienstleister gemeinsam ein Fraud-Detection-Modell trainieren, ohne ihre Trainingsdaten auszutauschen. Das gemeinsam trainierte Modell soll effektiver sein als Modelle, die nur mit den Daten eines einzelnen Unternehmens trainiert wurden.

Anwendungsfall #4 IoT (Secuvera):

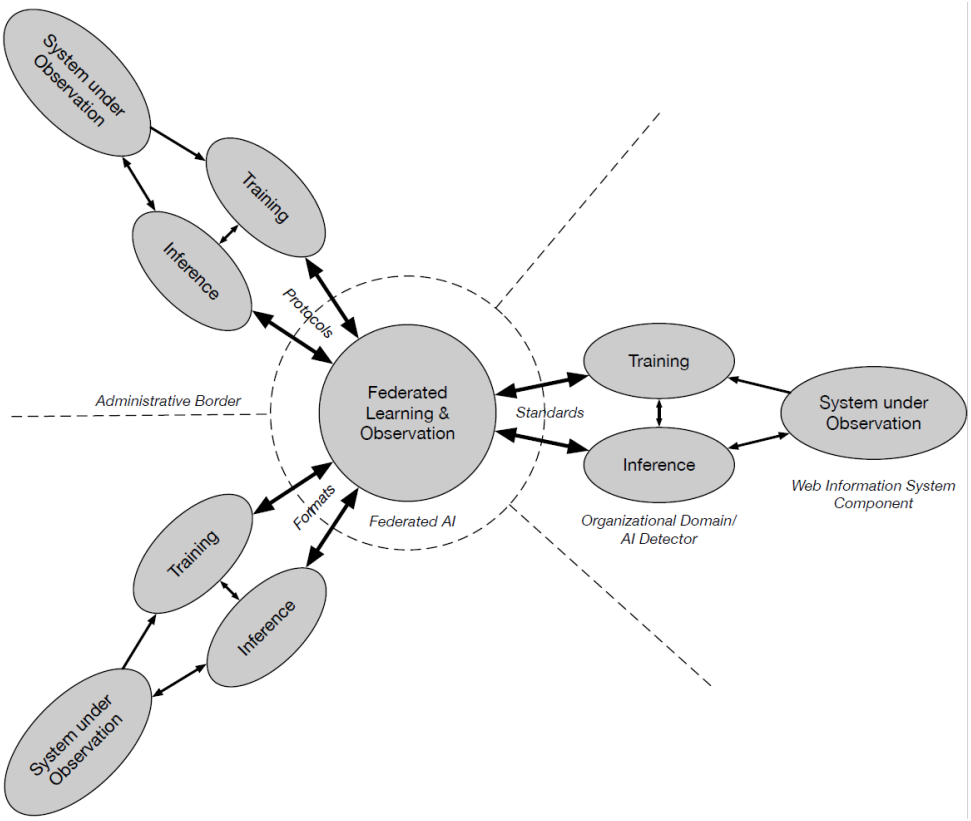
Der Use Case “Einschleusen/Manipulieren von Komponenten in Industrienetzen” befasst sich mit der Erkennung von Angriffen in Industrienetzen durch Analyse der Netzwerkkommunikation und Prozessdaten der OT-Komponenten. Ziel ist es, Manipulationen von Engineering Workstations und Industriespionage durch neue Komponenten zu erkennen. Der Fokus liegt auf der Föderierung von Netzwerk- und Prozessdaten, nicht auf der Föderierung zwischen verschiedenen Industrieanlagen.

Anwendungsfall #5 Angriffe auf Basis von Auth Protokollen (TUBS):

Dieser Use Case untersucht Angriffe auf WebApp-Authentifikation durch OAuth 2.0 und OIDC, wie sie von NetID verwendet werden. Ziel ist die Erkennung von Angriffen aufgrund von Fehlkonfigurationen oder fehlerhaften Protokollflüssen. Dazu werden Daten von allen beteiligten Protokoll-Endpunkten horizontal föderiert. Ein Monitor bewertet den aktuellen Flow in Echtzeit und gibt diese Vertrauensinformation vertikal an die nächste Partei, wie z.B. den Webseitenprovider, weiter. Bei niedriger Vertrauensstufe kann diese Information zur Erkennung und Verhinderung von Kreditkartenbetrug genutzt werden.

2.2 Architekturkonzept

Aus der kombinierten Betrachtung von Anwendungsfällen wurde ein Architekturkonzept abgeleitet. Die Architektur skizziert einen offenen, ganzheitlichen Systementwurf, auf dessen Basis Erzeugung, Einsatz und Verwaltung domänenspezifischer und auch föderierter KI-Modelle zur Erkennung und Behebung von Sicherheitsbedrohungen ermöglicht werden soll.



Schematische Darstellung der KIWI-Gesamtarchitektur (Stand Dez. 2020)

Die Architektur basiert auf der Annahme eines föderierten Web-basierten Informationssystems, dessen Komponenten sich über verschiedene technisch-administrativen Domänen erstrecken (Unternehmen, Abteilungen/Funktionsbereiche, Teams). Jede Domäne beinhaltet einen KI-Detektor für die Erkennung von und Reaktion auf sicherheitsrelevante Ereignisse, der auf einem lokalen ML-Lebenszyklus mitsamt Training und Inferenz basiert. Verschiedene Domänen sind mittels föderierter KI-Mechanismen verbunden, wobei Protokolle und Formate zum Einsatz kommen, für die eine Standardisierung erfolgt. KIWI fokussiert insbesondere föderierte Lernmechanismen in zentralisierter und dezentralisierter Form, aber auch der Austausch von erkannten sicherheitsrelevanten Ereignissen und die Reaktion darauf durch koordinierte Gegenmaßnahmen sind relevant.

3 Anonymisierter Datensatz für die Partner

Um die Kooperation um den Anwendungsfall ” Missbrauch von Identitäten” zu stärken wurde bei Mail&Media folgend ein Datensatz entwickelt und den Partnern zu Verfügung gestellt.

Die Ergebnisse von Machine-Learning Verfahren sind stark von den zugrundeliegenden Datenverteilungen abhängig. Ob sich ein entwickelter Ansatz auf andere Daten- oder Label-Verteilungen erfolgreich übertragen lässt sich i.A. nicht vorhersagen. Daher war es unser Ziel mit den Partnern (auch) mittels geteilter, realistischer Daten zu kooperieren.

Intensiven, wiederholte Konsultationen mit Datenschutzexperten in unserer Organisation machten jedoch deutlich, dass in Ermangelung expliziter Nutzerzustimmung (nicht repräsentativ) nur stark anonymisierte Daten Datenschutz-rechtlich vertretbar sind.
Um die Kooperation hier fortsetzen zu können wurde als Teil der Vorhaben AP 211 und 312 daher ein anonymisierter Datensatz erzeugt. Der Datensatz umfasste 700Tsd randomisiert gesampelte Login-Vorgänge der Marken web.de und GMX, mit folgenden Attributen:

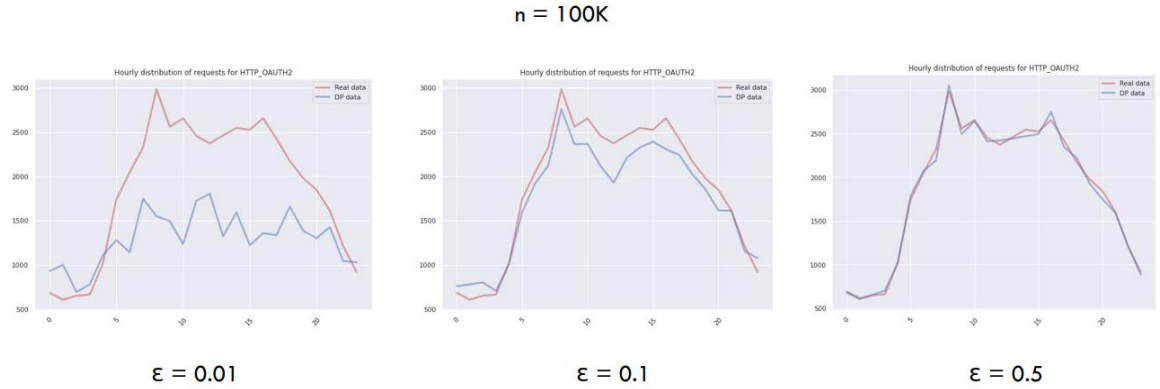
Attribut-Name	**Anonymisierungsmaßnahme**
HappenedTimestamp	
remoteIp	Spezifischste 16bit gelöscht – Ambiguität ist 2^16 Möglichkeiten.
GeoIp.country	
ExternalUserRequest.type	
UserAgent	
hashedCookies["__Host-ls.rec"]	Gehasht (SHA256). Cookie ist eine zufällige UUID ohne Personenbezug.
UiUserApplication.type	
ProvidedIdentifier.type	
providedIdentifier.value	Alles bis auf die erste Stelle gelöscht.
AccountIdentificationProcess.success	
accountIdentifier.sub	256bit Pepper hinzugefügt, dann SHA256 gehasht.

Die Erzeugung des Datensatzes erforderte wiederholte Abstimmungen mit Datenschutz- als auch Security-Beauftragen, das Ergebnis stieß aber nicht auf ein ausgeprägtes Interesse bei den Partnern, daher wurden die Kooperationsbemühungen dann vor allem in konzeptionellen und analytischen Bereichen u.A. durch die zeitweise Anstellung von Doktorand David Monschein fortgesetzt (siehe auch Kapitel 6).

4 Datenschutz-optimierte Synthese und Speicherung von Metadaten

Die in Kapitel 3 dargestellten “Datenschutzherausforderungen” zeigten sich auch Unternehmens-intern. Die Erkennung von Hackern geht oft auf Abweichungen von “normalem Verhalten” zurück, und damit implizit oder explizit auf den Vergleich mit “historischen” Daten. Mit den seit dem 25. Mai 2018 geltenden DSGVO-Richtlinien ist das Speichern von Nutzerdaten für einen längeren Zeitraum (ohne Zustimmung des Nutzers) nicht mehr zulässig. Nach den aktuellen Richtlinien ist es nur noch möglich, Daten maximal einen Monat lang zu speichern. In KIWI möchten wir jedoch mehr historische Daten speichern, damit unsere Modelle besser funktionieren oder historische Analysen durchgeführt werden können. In teilweiser Bearbeitung der Vorhaben AP 211,212,312 und 413 wurden Möglichkeiten untersucht, DSGVO-konform und "risikofrei" Daten langfristig zu speichern.

Eine Möglichkeit besteht darin, eine synthetische Version unserer Daten zu erstellen, die ähnliche Eigenschaften wie die Originaldaten aufweist. Wir verwenden Differential Privacy (DP) als Werkzeug, um "falsche" Daten für uns zu erstellen, die aber statistisch konsistent mit den tatsächlichen Datenwerten bleiben und zufällige Abweichungen ermöglichen. Der Grad der Abweichung (Anonymisierung) ist dabei über DP’s ε-Parameter wählbar.



Verteilungstreue stündlicher OAUTH Requests in Abhängigkeit von ε.

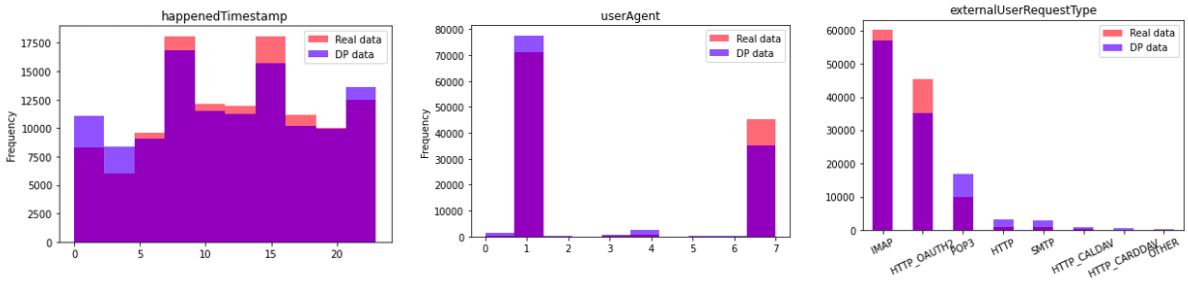
Um dieses Verfahren anwenden zu können, musste es jedoch bzgl. folgender Fragen weiterentwickelt werden:

- Wie fügt man Rauschen zu kategorischen Daten hinzu? Wie erhält man die Korrelation zwischen Attributen?
    - Grundidee ist ein kombiniertes Sampling mehrerer kategorischer Attribute zusammen – damit wird eine Joint Probability Funktion definierbar, die automatisch auch bei “unmöglichen” Attribut-wert-kombinationen “das Richtige” tut.

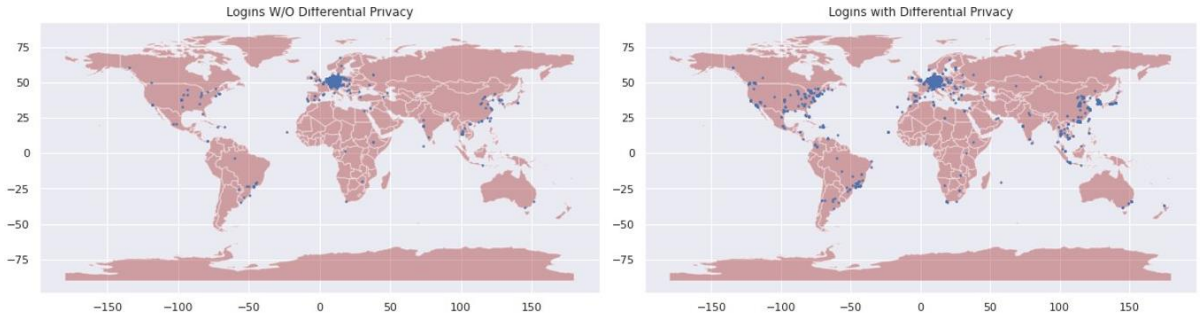
- Wie fügt man Rauschen zu String/Text Daten hinzu?
    - Indem Text zu Embeddings (z.B. FastText) transformiert wird. Optional können dann über Clustering-Verfahren den String-Embeddings Klassen bzw. Kategorische Werte zugewiesen werden.

- Wie misst man die Datenqualität der synthetischen Daten?

- Die Anwendung o.g. Erweiterungen auf komplette Datensätze wurden dann mit den Rohdaten-Verteilungen verglichen (s.u.).



Vergleich der Verteilungstreue kombinierter Zahlen-Text-Kategorie-Daten für ε= 0.1, n = 100K.



Vergleich der Verteilungstreue von geo-spatial Verteilungen. Nutzerschwerpunkte bleiben erhalten. Bei schwach repräsentierten Regionen ist die erzeugte Verteilung weniger getreu – was aber aus Daten-schutz-Sicht erwünscht ist.

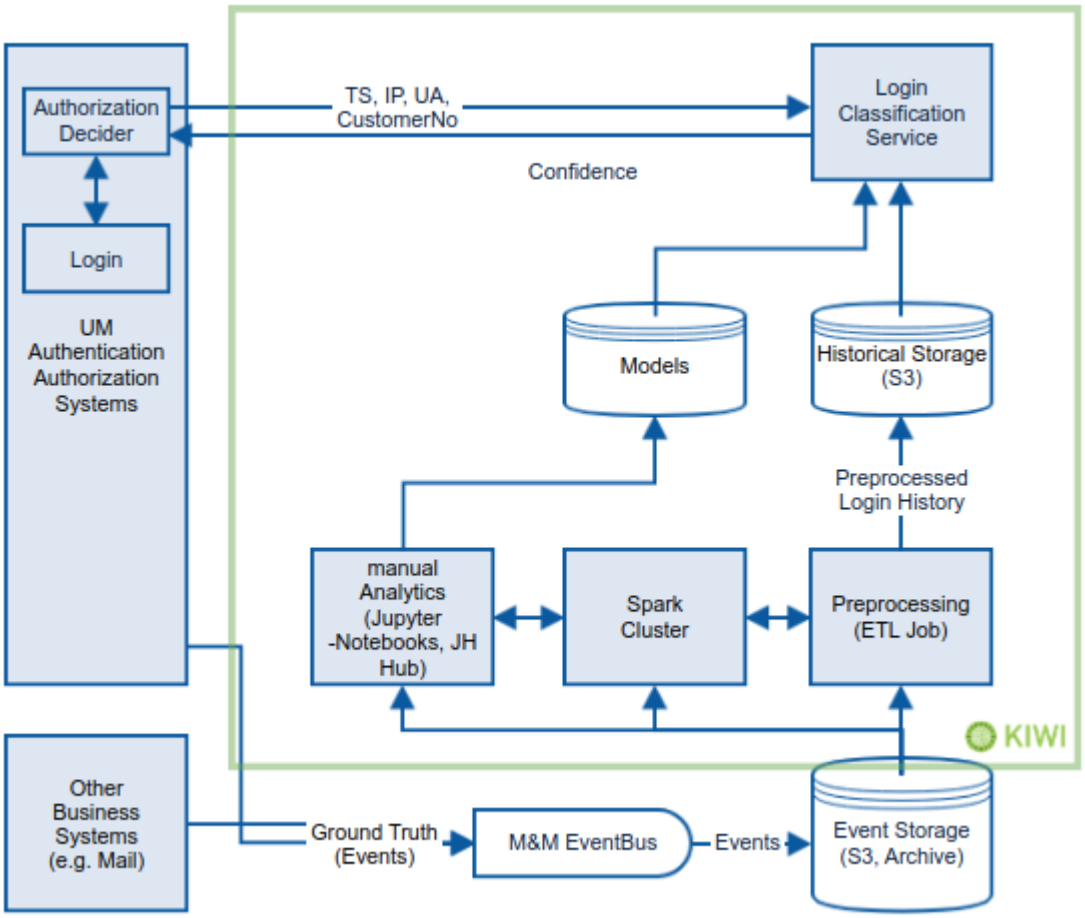
Das Verfahren, mit unseren Weiterentwicklungen erreicht die o.g. Ziele. Es ist jedoch teuer und kann grundlegende Zielkonflikte nicht auflösen:

- Je stärker (via ε) anonymisiert wird, desto schlechter "performen" darauf trainierte Modelle auf Echtdaten. Dies wird durch den 2. Zielkonflikt besonders ausgeprägt.
- Die fachliches Zielstellung unser zu entwickelnden KI-Systeme ist in der Regel eine identifizierende ("Ist das der legitime/bekannte Nutzer?").

Aus diesen Gründen wird eine breite, produktive Verwendung aktuell nicht geplant.

5 Arbeiten zu Daten- und Analytik- Infrastruktur

Mit der für KIWI beschafften Server (siehe Kapitel 9) wurde, unter Hilfe von Datacenter-, Networks- und Operationskollegen, ein Spark Analyse- und Trainings-cluster aufgebaut.



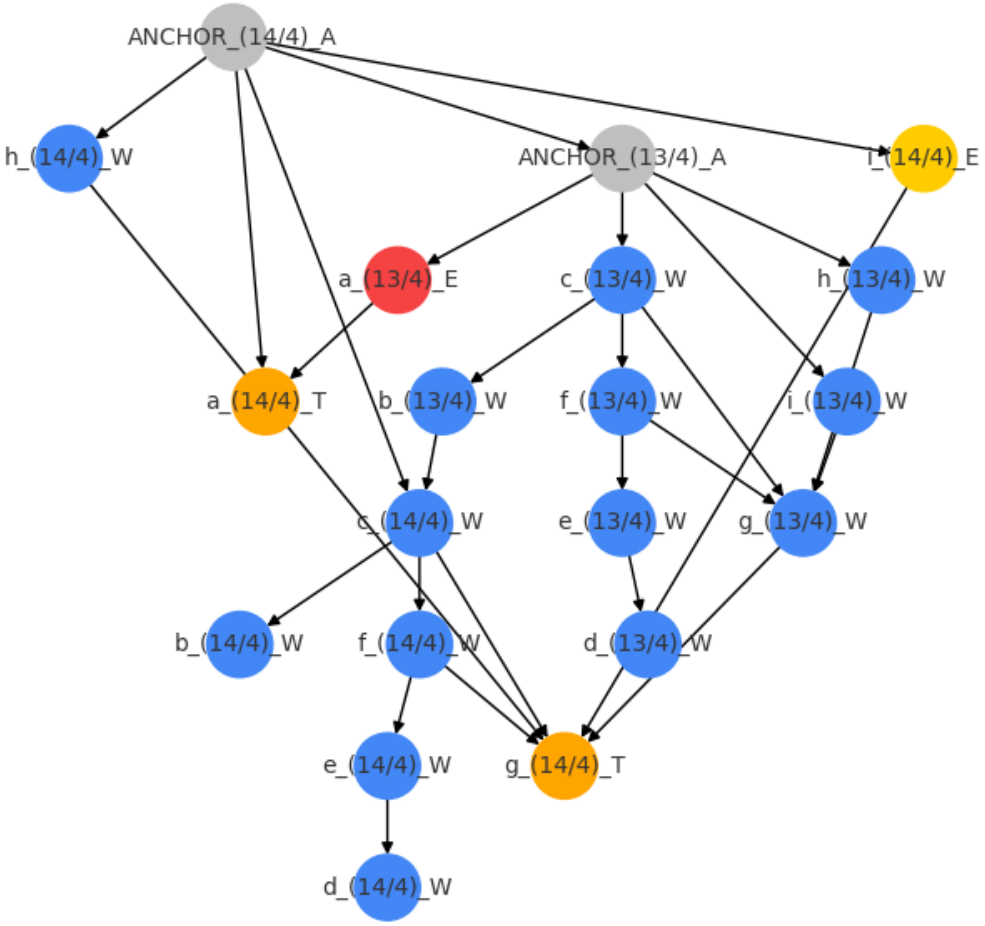
Integration der KIWI-Komponenten (grün) in bestehende Infrastruktur. Alle Systeme sind horizontal skalierbar, bei einem aktuellen Eventaufkommen von 1-2Mrd. Event/Tag. Analysen und KI-Modellentwicklung umfassen typisch 30-100Tage Eventdaten plus vor-aggregierte Daten.

Lasttests nötigten wiederholt zu aufwendiger Fehlersuche und Justierungen der Konfigurationen über fachliche Teams hinweg; so z.B. wegen der Kannibalisierung von Netzwerkkapazitäten durch die vom Cluster verursachten hohen Datenaufkommen.

Aufgrund von DSGVO-Vorgaben, aber auch aus Kostengründen kann nicht längerfristig mit Rohdaten gearbeitet werden, ältere Daten werden daher sukzessive immer stärker anonymisiert (e.g. siehe Kapitel 4) und aggregiert. Nach Tests mit einem Stream-Processing-Framework (Apache Flink) entschieden wir den ETL-Job als Batch-Job auf Basis von Spark zu implementieren. Dies bietet, bei dem Nachteil der Verarbeitungsverzögerung, folgende Vorteile:

- Transformations- & Filter-Code ist dem Analyse-Code, der in Jupyter Notebooks entsteht sehr nahe bzw. sehr ähnlich.
- Effizienz: Columnar-Storage File-Formate wie Apache Parquet zeigen für die hier typischen Daten erhebliche Kompressionsraten.

- Einfacherer Umgang und Sicherstellung der Korrektheit bei Fehlern, zeitweisen Ausfällen, Retry's und Abhängigkeiten zwischen Transformationen.



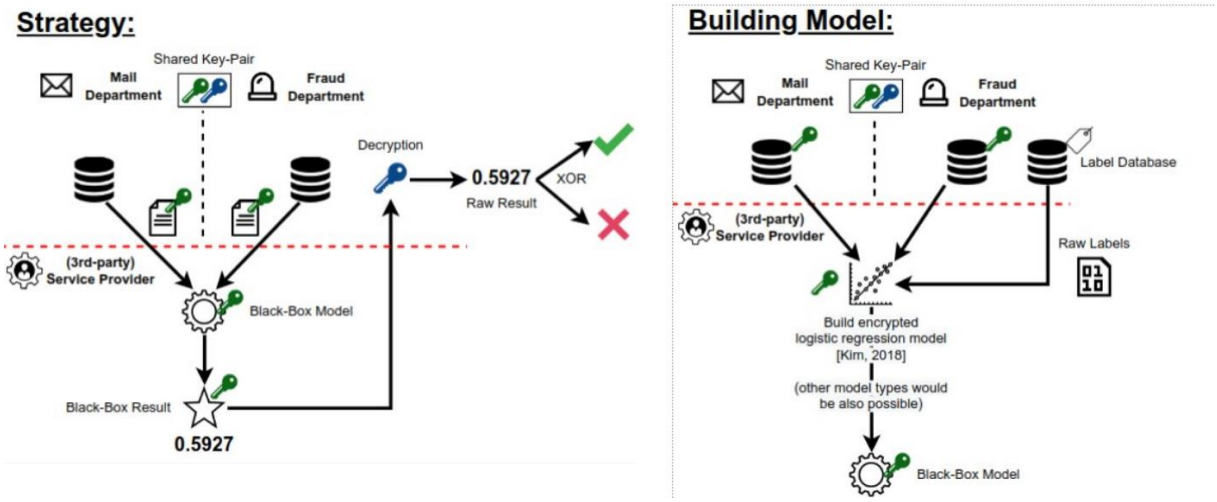
Transformationsgraph des KIWI ETL Jobs vom 13. bzw. 14.April. Mit einer fehlgeschlagenen Transformation (rot) und davon abhängigen/betroffenen Folgetransformationen (orange). Die aktuelle Konfiguration berechnet 20, voneinander abhängige Transformationen auf >10^9Events/Tag und ist "selbstheilend" bzgl. temporär fehlschlagender Transformationen.

6 Betrugserkennung mit föderierten KI-Modellen in Kooperation mit der 1&1 Telecommunications SE

Mit der 1&1 Telecommunications SE wurde für den Anwendungsfall #3 (Fraud Detection) eine Kooperation verfolgt. Motiviert durch den mit GMX bzw. Web.de Accounts bei der Schwestergesellschaft durchgeführten Betrug (z.B. betrügerische Mobilgerätbestellung) sollte die Frage untersucht werden, ob mit föderierten KI-Modellen und –Prozessen dieser DSGVO-gerecht reduziert werden könnte. Bearbeitet wurden dabei Fragenstellungen der Vorhaben AP 214, 216, 312, 314, 411 und 514.

Die Rahmenbedingungen für die Kooperation sind denkbar gut: Beide Seiten (auch monetär) stark motiviert, die Frage nach Geschäftsgeheimnissen ist innerhalb der 1&1 Gruppe weniger relevant. Jedoch stellte sich auch hier bald die Frage nach der "Datenschutzrechtlichen Bewertung einer erweiterten Betrugsprüfung" (Gutachten der Kanzlei Loschelder). Diese Diskussionen zogen sich im Projektverlauf über ein Jahr hin, und verzögerten und erschwerten das Vorhaben erheblich.

Unterdessen wurden in Kooperation mit der HKA (Prof.Waldhorst / D.Monschein) verschiedene Konzepte entworfen, um die Kooperation der Fraud-KI-Systeme möglichst DSGVO-verträglich zu gestalten.



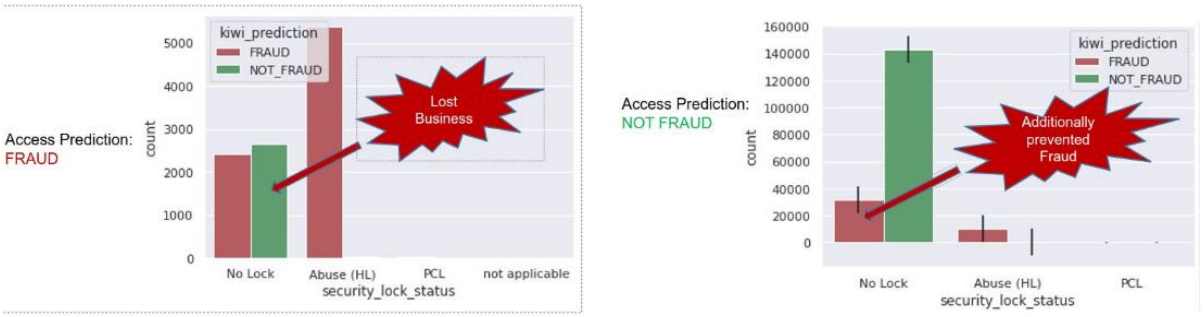
DSGVO-freundliches Abstimmungsverfahren für föderierte KI-Systeme auf Basis homomorpher Verschlüsslung (D.Monschein/HKA). Die homomorpher Verschlüsselung eigene hohen Berechnungskosten und Verzögerung wäre in diesem Anwendungsfall verkraftbar.

Die parallel fortgeführte Entwicklung eines KIWI/Mail&Media Fraud-Modells erwies sich dagegen als höchst erfolgreich: Die in einem abschließenden Blind-Test erzielten Fraud-Erkennungsraten lassen bei produktiver Umsetzung auf eine Reduzierung des jährlichen Betrugsschadens bei der 1&1 Telecommunications SE um mehrere Millionen Euro hoffen. Verhandlungen auf Vorstandsebene dazu stehen an.

7 Betrugserkennung innerhalb der Mail&Media GmbH mit föderierten KI-Modellen

Innerhalb der Mail&Media GmbH wurde Anwendungsfall #3 (Fraud Detection) die Möglichkeit eines föderierten KI-Systems evaluiert (adressiert Vorhaben AP 216, 412, 514). Dabei wurden Modelle aus verschiedenen Unternehmensteilen verglichen bzw. kombiniert (“Ensemble”), darunter ein im Rahmen von KIWI in einer Bachelorarbeit trainiertes Modell [1]. Die KIWI-entwickelten Modelle arbeiten auf Features verschiedener Domänen (Mail, Usermanagement, Payment), und lassen im Ergebnis eine Verbesserung der Fraud-Erkennung erwarten.

Auf Details wird hier nicht eingegangen (Unternehmensvorgabe).



Vergleich des von KIWI trainierten Modells mit einem existierenden Anti-Fraud KI-System.

Eine Weiterentwicklung als Produktivsystem wurde erwogen, jedoch zunächst zugunsten des im Folgenden (Kapitel 8) vorgestellten Systems zurückgestellt.

8 Prototypisches KI-System zur Erkennung von Identitätsmissbrauch

Auf Basis der Vorarbeiten zur längerfristigen DSGVO-konformen Synthese und Speicherung von Meta-Daten und zur Analytik Infrastruktur, sowie der gewonnen Erfahrungen der vorangegangenen Anti-Fraud Experiments bzw. Kollaborationen wurde folgend ein **KI-System entwickelt, welches schwache Authentifikation (e.g. durch ein Nutzer-gewähltes Passwort) auf Basis von Metadaten transparent verstärkt**. Es soll erstmals zur synchronen Klassifizierung in einem kritischen Business-Prozess (Login) nutzbringend einsetzbar sein. Daraus resultiert ein breites Spektrum von Anforderungen (deren Beschreibung den Rahmen dieses Dokuments sprengen würde) erfüllen. Entsprechend werden eine Reihe von Vorhaben AP adressiert: 112, 113, 114, 212, 215, 216, 311, 312, 313, 315, 411, 412, 414, 511, 512, 513, 514.

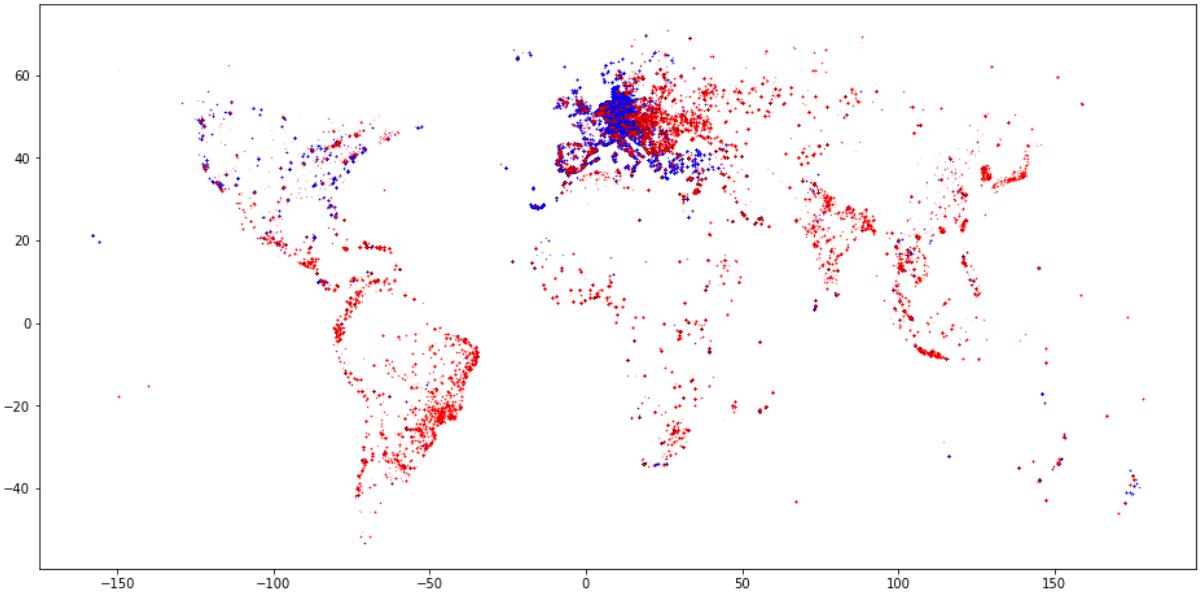
Neben technischen Herausforderungen (z.B. Hochverfügbarkeit, Skalierbarkeit bis zu einigen Tausend Klassifizierungen pro Sekunde, Antwortlatenz <20ms, p99) waren fachlich u.A. folgende Herausforderungen zu adressieren:

- Fachliches Kernziel ist die Klassifikation legitimer Accountnutzer und –nutzung als Solches anhand von Metadaten.
- Es muss von einem (geringeren) Anteil schon kompromittierter Identitäten/Accounts ausgegangen werden. Wie erreicht das System eine Reduktion dieser?
- Ziel ist eine Verbesserung gegenüber der existierenden einfachen Passwortauthentifizerung zu geringen “Convenience”-Kosten der Nutzer. Diese erreicht aber schon eine hohe Präzision (>95%).
- Für klassische “supervised learning” Ansätze fehlen ausreichend qualitativ hochwertige Label:
    - Hacker bestätigen nicht “Ja, erwischt!”.
    - Die, per Annahme vertrauenswürdigen, Label aus Mehrfaktorauthentifizierungen sind sehr wenige, und “einseitig”, d.h. ein Misserfolg ist kein Beweis eines Hack-Versuchs.
    - Bisherige Anti-Abuse-Prozesse im Unternehmen bewerten und sanktionieren den Account – ohne eine Erkennung oder gar Trennung verschiedener Akteure im Account (Hacker und Nutzer).
    - Diese, für die Domäne charakteristischen Probleme wurden in [2] eingehend untersucht, und sind auch zukünftig Arbeits- und Forschungsgegenstand.

Die folgenden Experimente wurden durchgeführt, um Referenz- bzw. Erwartungswerte für ein mögliches Produktivsystem zu erhalten.

8.1 Experiment: Lokales Modell

Auf Basis eines gut erkennbaren Angriffs gegen OAuth2’s Resource Owner Password Credential Grant API und Kunden Logins mit starker Authentifizierung (OAuth2 Refresh Token Grant) wurde ein Datensatz mit rund 1Million Records erstellt.



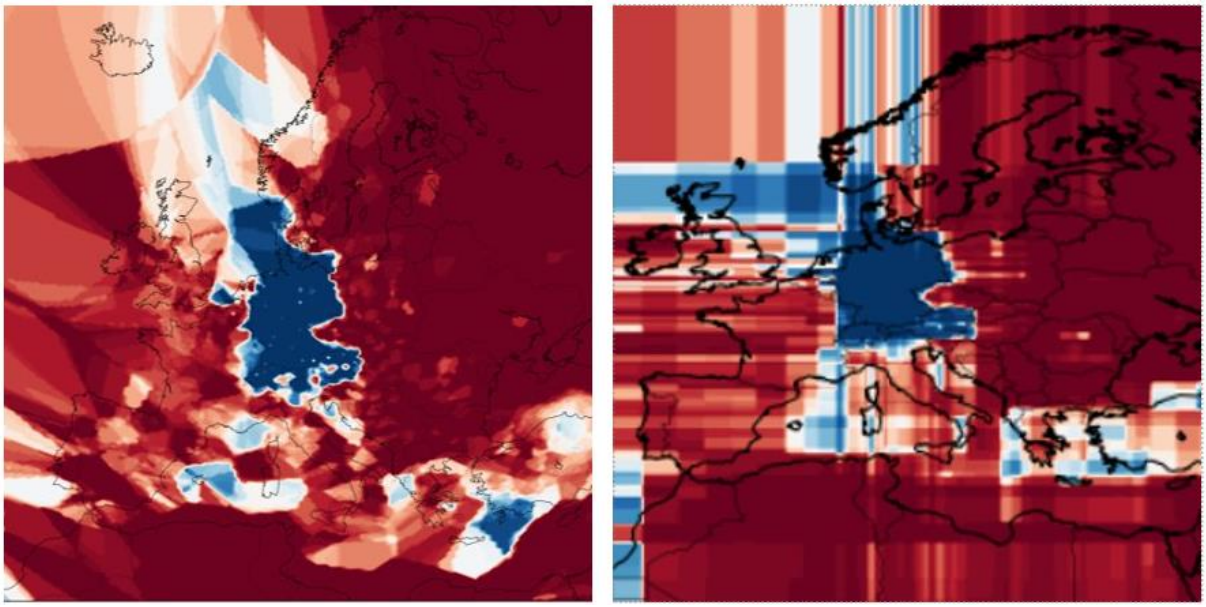
Geospatialer Plot von Hacker Logins (rot) und Kunden (blau).

Für jeden Account (“local model”) wurden nun mit verschiedenen Verfahren (Logistic Regression, k-NearestNeighbor, Support-Vector-Machine, Cache) Modelle trainiert, getuned und verglichen. Z.T. mit guten Ergebnissen (AUC bis zu 0.986), jedoch wurden inhärente Nachteile offensichtlich:

- Je nach Accountnutzung und DSGVO-Vorgaben zur Speicherdauer liegen z.T. nur unzureichende Feature Daten vor.
- Es kann kein Kontext- bzw. Populationswissen erlernt werden (“Nutzer, die sich zur Urlaubszeit vom Flughafen einloggen, tauchen danach gern aus Mallorca auf.”)
- Für eine synchrone Einbindung in den Login-Prozess wäre es erforderlich, innerhalb von Millisekunden unter Millionen von Modellen auf das jeweilige zuzugreifen und es auszuwerten – eine, unter Umständen sehr teure Skalierungsanforderung.

8.2 Experiment: Globales Modell

Auf dem gleichen Datensatz wurden wie zuvor wurden nun Verfahren bzw. Modelle für die gesamte Account- bzw. Login-“Population” trainiert.

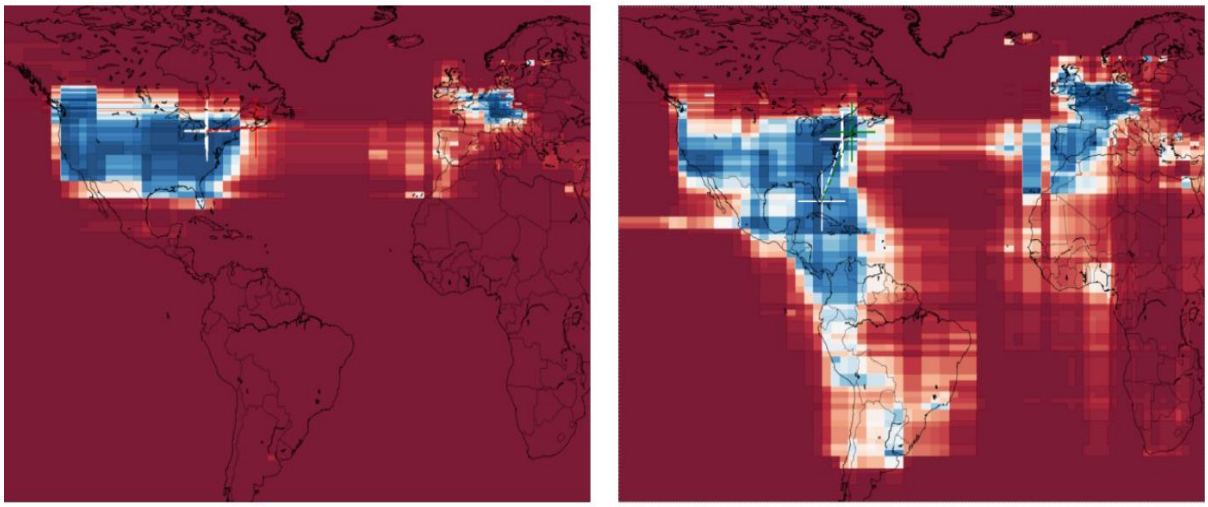


Geospatial-Plot der Entscheidungsgrenzen von kNN-Modell (links) und Boosted Tree Ensemble (rechts, XGBoost). Die Grenzen der D-A-CH Region sowie typische Nutzer-Urlaubsdestinationen (Gardasee, Mallorca) wurden als vertrauenswürdig (blau) erlernt.

Der Ansatz eines (Populations-)globalen Modells vermeidet o.g. Nachteile Account-individueller Modelle weitgehend, kann in dieser Form jedoch nur beschränkt individuelles Verhalten von Accounts modellieren.

8.3 Experiment: Hybrides Modell

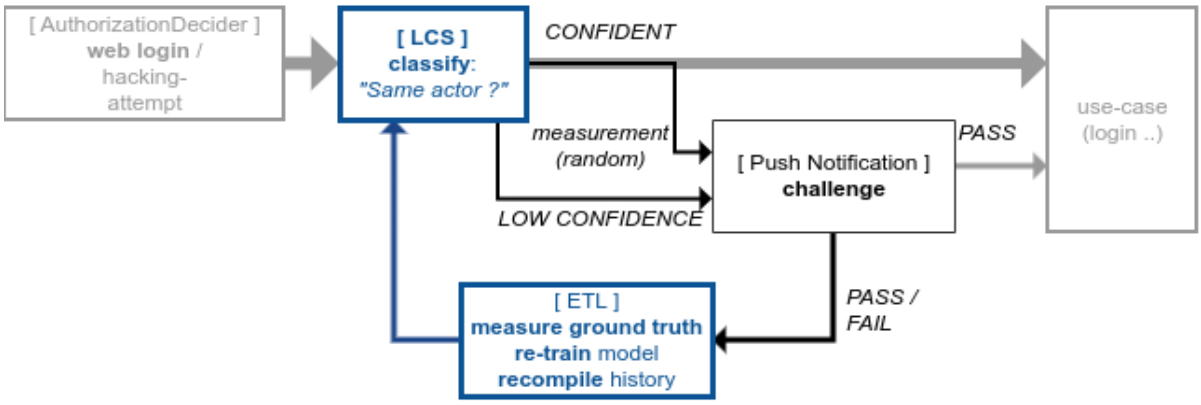
Können die Vorteile o.g. Ansätze in Einem vereint werden? Um das herauszufinden, wurde, erneut auf den gleichen Daten, ein “hybrid” Modell entwickelt: Ein einziges trainiertes Modell wird mit der Historie eines echten Kunden und einem zusätzlichen Login konfrontiert. Die Aufgabe des Modells besteht darin zu bestimmen, ob der jüngste Login in die Historie des Kunden passt oder nicht (legitim oder Angreifer).



Geospatial-Plot der Vertrauensregionen (blau) zweier Accounts im Vergleich: Links mit Loginhistorie aus dem Nordosten der USA, rechts ähnlich, jedoch zusätzlich mit einem Login aus dem, z.T. Spanisch-sprachigen Florida.

Die Ergebnisse des Experiments sahen erfolgversprechend aus, und potenziell produktiv einsetzbar und weiterentwickelbar, und so wurde entschieden, auf dieser Basis ein prototypisches Testsystem für die Live-Umgebung zu entwickeln.

8.4 Konzept der Prozessintegration



Business-Prozess-Integration und automatischer Training-Feedback-Loop. Qualitätsmessung durch kontinuierliches Auswerten von Nutzerantworten randomisierter Push-Notifikationen (schwarz). KIWI Systeme Login-Classification-Service ”LCS” und Extract-Transform-Load Job ”ETL” (blau).

Feedback- & Retrainingszyklus:

1. Klassifikation bei jeder Gelegenheit zum Hacken (Login). Die Semantik der Klassifizierung ist die der kontinuierlichen Kontrolle (“Gleicher Akteur wie bisher akzeptiert?”).
2. Vom Nutzer werden bei (niedrigem) Vertrauen, als auch, mit geringer Wahrscheinlichkeit, zufällig, erweiterte Authentifizierungsbeweise gefordert (z.B. per Push-Nachricht “Ist dies ihr Login?” auf das Mobilgerät). Die Ergebnisse stellen nach Annahme eine Grundwahrheit dar.
3. Auf Basis dieser neuen Nutzerantworten wird das Modell automatisiert nachtrainiert.

8.5 Arbeiten zur Infrastruktur, Risikomitigation

Softwaresysteme im Live-Betrieb unterliegen bei 1&1 einer Reihe von Anforderungen (Code-Qualität, Test-Abdeckung, Security-Prozess) auf die hier aber nicht näher eingegangen werden soll, da sie den Rahmen sprengen würden und nicht spezifisch für dieses System sind. Speziell ist jedoch der Einsatz eines trainierten KI-Systems in einem kritischen Business- bzw. Kundenprozess mit stringenten Availabilty-Anforderungen; die Frage z.B. nach dem Risiko, dass ein lernendes System “über’s Wochenende” vielen Kunden fälschlicherweise den Zugriff auf ihren Accounts verweigern könnte, lässt Verantwortliche verständlicherweise sehr schnell sehr vorsichtig werden. Entsprechend mussten LCS und ETL-Job in produktive Monitoring- & Alerting-Systeme integriert werden. Der Aufruf des LCS aus dem Login- bzw. Autorisierungsprozess heraus wurde über Monate langsam gesteigert und Kundenfeedback währenddessen intensiv beobachtet und analysiert. Neue Modelle werden zunächst silient/passiv deployed, d.h. befragt, aber in der Login-Entscheidung nicht berücksichtigt. Kapitel 5 beschreibt die technischen Arbeiten der Integration genauer.

8.6 Ergebnisse des Testsystem in Live

Seit Anfang 2024 wird das System testweise für alle Web-Logins aufgerufen. Mittels weniger, randomisiert-ausgespielter Push-Nachrichten werden Klassifizierungsfehler beidseitig messbar, mit folgendem Ergebnis:

- False Negative Rate 1.14%
- False Positive Rate 1.15%
- F1-Score 98.9%
- Area Under Curve 0.998

Das Monitoringsystem zeigt zudem Evidenz einer starken Korrelation der Modellaussagen zu Kundenantworten.

Screeshots aus dem Monitoringsystem des LCS Live-Systems vom 24.1.24 (links) und 16.2.24 (rechts). Jeweils oberes Panel (rot): Anteil von ”low confidence” Modell Antworten an Web Logins Requests. Jeweils unteres Panel: Kundenreaktionen ”Deny” (rot) auf eine Push-Nachricht ”Ist das ihr Web-Login?”. Eine klare Korrelation ist erkennbar., d.h. der Angriff wurde gut erkannt. Dagegen bleibt die Rate der ”Confirm” Antworten (gruen) im Wesentlichen unbeeinflusst, was eine gute Präzision vermuten lässt.

In der Summe werden **Woche für Woche einige Tausend Hackversuche erkannt, und als solche von den Nutzern bestätigt**. Da LCS als letztes Glied in einer Reihe von Checks eingebunden ist, sind dies anderweitig nicht erkannte Hackversuche von Kundenaccounts, die ohne das System Erfolg gehabt hätten.

8.7 Konkrete Planungen

Aufgrund der guten Resultate im Live-Test, sowie der beobachteten, anhaltend hohen Angriffsintensität führt 1&1 die Entwicklung des Systems fort und erwägt ein verbessertes System längerfristig zur Reife zu bringen, auch zum Einsatz bei anderen Use-Cases, z.B. Accountdaten-Änderungen, oder Mail- bzw. Mobil-Logins. Vorbedingung sind allerdings umfangreiche Überarbeitungen bzw. Neuentwurf und -implementierung, u.A. der Effizienz, der Lastfähigkeit, der Feature-Storage im LCS, oder der Erstellung von Trainingsdaten.

9 Ergänzende Angaben

9.1 Wichtigste Positionen des zahlenmäßigen Nachweises

Die wichtigsten unmittelbaren Vorhabenkosten waren (Gesamtnachkalkulation, Euro):

- Personalkosten 1.254.135,-
- Abschreibungen auf vorhaben-spezifische Anlagen 52.128,-
- Reisekosten 6.363,-

Die unmittelbaren Vorhabenkosten entsprachen weitgehend der Vorkalkulation, und lagen in der Summe (Verwendungsnachweise, Position 0881, “gesamte Selbstkosten des Vorhabens”) leicht unter der ursprünglichen Vorhabenkalkulation.

9.2 Verwendung der Zuwendung

Die Zuwendung der unmittelbaren Vorhabenkosten wurde im Einzelnen wie folgt verwendet:

- Personalkosten wurden entsprechend der Vorhabenplanung für die Besetzung von 4 Analysten- bzw. Entwicklerstellen aufgewendet.
  - Die Kollegen haben an den Arbeitspaketen kollaborativ und überlappend gearbeitet. Dementsprechend findet eine weitere Aufschlüsslung der Zuwendung auf Stellen oder deren Arbeitspaketleistungen nicht statt.

- Abschreibungen auf vorhaben-spezifische Anlagen wurden entsprechend der Vorhabenplanung für Vorhaben-spezifische Server verwendet:
  - Anlässlich der “Arbeiten zu Daten- und Analytik- Infrastruktur" wurden Mitte 2021 12 Dell Server sowie 12 GPUs gekauft und installiert.
  - Screenshot Serverbestellung:

Cost Center: 1041_45613 (KIWI)

SAP Order Number: 4500110561

Order Details: 12x Systems a (price: 8020.08):

SAPID	Name	Amount
None	R640v2	1
607945	2x6230R Gold(a 26Cores, 2.1GHz)	1
607321	SSD 480GB ATA MU S4610 2,5 512e	2
607812	DELL RAM 32GB RDIMM 3200MTs	12
606840	X710 DP10GB DA/SFP+ I350 DP NDC	1
606510	Dell HW PERC HBA330 12G Minicard	1
606811	PSU Single (1+0) 1100W	2

Order 1 10G FSP+ cables

- Screenshot GPU Bestellung:

  - Um die Anforderungen genauer spezifizieren zu können, fand die Bestellung später statt als im Vorhaben veranschlagt. Dies hatte eine geringere Abschreibungsdauer zur Folge. Eine anfänglich vorgesehene Erweiterung der Hardware im Projektverlauf konnte aufgrund ausreichender Kapazitäten und teilweiser Nutzung von Haus-interner Cloud-Infrastruktur komplett entfallen und führte zu deutlichen Einsparungen gegenüber der Vorhabenplanung.

- Reisekosten wurden für regelmäßige, etwa halbjährliche Workshops mit den Partnern aufgewendet. Aufgrund der Mehrheit der Partner fanden diese in Karlsruhe statt, und bedingten Anreise und jeweils eine Hotel-Übernachtung unserer, in München angestellten Kollegen.
  - Die Reisetätigkeit (Corona, keine internationalen Reisen) und damit auch Kosten fielen deutlich geringer aus als in der Vorhabenplanung erwartet.

9.3 Veröffentlichungen

[1] M.Khazma (2022): "Verwendung von maschinellem Lernen zur Klassifizierung missbräuchlicher Registrierungen”

[2] H.Bader (2024): “A Reconstruction-Based One-Class Approach to Login Classification”