

# AI Team Progress 2021-2022

Vorwarnung: Viel Arbeit im AI Team bezog sich auf Verbesserung der Suche, da es als "Core Produkt" angesehen wird, statt wie Oscar ständig an neuen MVP's zu arbeiten die aber selten in einen ausgereiften Zustand kommen.

## 2021

### Job2Jobs und Job2Skills (Semantische Suche)

- Verbesserung der Modelle im Bezug auf weibliche sowie Gender-neutrale Begriffe (z.B. Sales Mangerin, oder Sales Manager/-in)
- Erarbeiten einer "Zielstellung" für Job2Jobs um die Performance unserer Modelle zu vergleichen. (Nicht vollständig abgeschlossen, da abhängig von anderen Teams mit Recruiting Knowhow)
- Anfang von kontinuierlichen Arbeiten an einer Verbesserung des Models:
  - Systematisches Untersuchen der Model-Parameter
  - Verbessern der Trainings Daten mit "Normalisierung"-Algorithmen (zieht sich durch großteile von 2022)

### Smart Search V2

- Konzipierung von nächsten Schritten um die Smart Search zu verbessern:
  - Erfassen von verschiedenen Feldern auf den verschiedenen Netzwerken sowie deren Relevanz für eine Suche nach Job Titeln und Skills (Mit Hilfe von Laura und Marcel)

### Project2Search, oder auch Profile Recomendations

- Erarbeiten und Umsetzen einer alternativen Suche auf Basis von Projekt Daten.
- Insbesondere: Implementieren eines Mechanismus für direktes Feedback (Kandidat passt, oder passt nicht)

- Tracken dieser Daten als Grundlage für zukünftige datengetriebene Entscheidungen (zuerst BI, irgendwann Testing der Machine Learning Modelle, eventuell irgendwann als Trainingsdaten, je nachdem welche Menge an Daten wir bekommen)

## **Profile Matching (Andere Profile des gleichen Kandidaten auf verschiedenen Netzwerken)**

- Konzipierung und Entwicklung eines ersten Prototypen für ein Machine Learning Model das lernt ob zwei Profile dem selben Kandidaten gehören, oder nicht.
  - Beschaffen von Trainingsdaten
  - Identifizieren von korrekten Performance Metriken
  - Erstellen eines ersten Test und Validation Sets
  - Trainieren eines ersten MVP
- Verbessern des “Could also be service” mit Hilfe dieses Profile Matching Models
- Erste Anfänge von “Profile Image Matching” (zieht sich durch großteile von 2022)

## **Interne Verbesserung:**

- Komplette Überarbeitung des Codes für die Suchlogik aus der Smartsearch, Project2Search und für interne Nutzung (zB generieren von Trainingsdaten) → Vorbereitung und Erleichterung der kontinuierlichen Verbesserung der Suchlogik (Zieht sich durch große Teile von 2022!!!)
- Verbesserung und Vereinheitlichung des Service Deployments (z.B. Migration der Enrichment Services in ECS)
- Erste Verbesserung des damaligen Geolocation Service und entwickeln eines automatischen Reportings wenn gefundene Koordinaten als “unsicher” eingestuft werden.
- Ausprobieren von “Topic Model” Ansätzen zum Klassifizieren von Skills nach Domänen wie “Tech”, “Business”, ... Größtenteils erfolglos.

# **2022**

## **Smart Search V2**

- Umsetzung einer neuen adaptiven Suchlogik. Das heißt, die gesuchten Job und Skill Terme werden auf den verschiedenen Netzwerken unterschiedlich gewichtet und gesucht. Kern Verbesserung: Weniger prominente Netzwerke können nun in der Smart Search gefunden werden.
- Kontinuierliche Verbesserung der Performance durch migrieren der komplexen Suchlogik in die Datenstrukturen auf Elasticsearch (zusammen mit Sebastian)
- Experiment: Relative Gewichtung der Netzwerke zueinander
  - Versuch eine konstante relative Gewichtung als Kompromis zu identifizieren → Gescheitert, existiert nicht
  - Möglichkeitsprüfung für Machine Learning Ansätze, welche eine Gewichtung je nach Job Titel vorschlagen kann → Blocked durch Mangel an Trainingsdaten (braucht weitere Untersuchungen)
- Entwickeln und Implementieren von Tests für die Smart Search für leichteres Testen durch QA und weniger Ausfälle durch fehlerhafte Queries.

## Smart Search V3

- Erstes Brainstorming und Konzipierung von nächsten Schritten für die Smart Search als Power Tool

## Keyword Extraction (JobAd2Search für besser Project Recommendations)

- Möglichkeitsprüfung für verschiedene Ansätze für das Parsen von Job Titeln, Skills, Locations und Companies von JobAds → Erfordert moderenes Machine Learning Model
- Konzipieren und Trainieren eines ersten MVP für Keyword Extraction
  - Labeln von Trainings und Test Daten
  - Trainieren und Testen erster Modelle zur Abschätzung von notwendigen nächsten Schritten (derzeit unzureichende Performance)
- Ausprobieren eines Mehrstufigen Ansatz für das vorsortieren von Sätzen und Absätzen

- Identifizieren eines eigenen “Recruiting-spezifischen” Sprachmodels als Grundlage für einen Erfolgreichen Keyword Extractor
  - Trainieren eines ersten MVP für ein Sprachmodel mit zusätzlichen Recruiting spezifischen Termen (z.B. Programmiersprachen)

## **Super Profile (als verbesserung des “Could also be service”)**

- Formulieren der Vision: “Suchen nach einem Kandidaten, statt nach einem einzelnen Profil des Kandidaten”, “Datenschutz konforme zusammenführung von allen Profilen des Kandidaten als echter Mehrwert für die Nutzer”
- Konzipierung und Begleitung der Umsetzung eines Super Profiles durch die Scraper. Hier erster Schritt: Verknüpfen von Profilen die sich gegenseitig verlinken.
- Vorbereitung und Verbesserung des Profile Matching Models in Vorbereitung auf das Super Profil

## **Profile Matching (Could also be Service) und Profile Image Similarity**

- Verbesserung des Models:
  - Trainieren eines Features auf Grundlage der Ähnlichkeit des Profilbildes (erste bessere Modelle trainiert, deployment in Produktion ausstehend)
  - Erarbeiten und implementieren von einfachen Algorithmen für das Anzeigen von Profile auch wenn sich das Model unsicher ist (auf Anfrage von CSM)

## **Job2Jobs und Job2Skills (Job Normalisierung als Fokus)**

- Verbesserung der Job Normalisierungs Pipeline:
  - Typos finden
  - Job und Skills aufspalten und zusammenfügen (z.B. Salesmanager → Sales Manager/-in)
  - Filtern von Orten und Company Names
  - Filtern von Termen mit geringer Statistik (z.B. Namen, Dauer der Stelle und ähnliches)

- ...
- Entwickeln einer ersten Erfolgsmetrik für das Job2Jobs Model auf Grundlage des unvollständigen Test-Datensatzes → Derzeit unter Revision
- Experiment: Vereinheitliches Training auf Jobs und Skills für das Verbessern von Ähnlichkeit zwischen deutschen und englischen Job Titeln sowie anderen "disconnected clusters" von Job Titeln. → Ergebnisse durchwachsen; Blocked durch fehlende Erfolgsmetrik/Test-Datensatz
- Erfassen und Eingliedern anderer verfügbarer Datenquellen für Job2Jobs Training, insbesondere gekaufte Daten von Oxyleads
- Fundamental neue Idee für Job Normalisierung (alter Ansatz nicht zufriedenstellend)
  - Definieren einer "Ziel-Datenbank" aus validen Job Titeln als Prototyp
  - Entwickeln einer Suche, oder Algorithmus, Job Titel aus den Profilen zu den klar definierten Job Titeln zu matchen. (WIP)

## Tracking Pipeline und besseres BI mit Zoho Analytics

- Reparieren und Instandhaltung der Daten-Pipeline, welche Tracking Events für Zoho Analytics aufbereitet.
- Erarbeiten eines Dashboards für CSM/Account zum Analysieren der Tracking Events im Bezug auf Nutzer und Firmen
- Kontinuierliche Ausbesserung und Zuarbeit für CSM
- Onboarding unseres neuen Data Analyst für zukünftigen Ausbau unserer BI Analysen.

## Interne Verbesserungen:

- Migrieren von den meisten Services zu einem "internen Service" ohne Authentifizierung für eine Vereinheitlichte Service Struktur
- Übernehmen und "Nachbauen" der Boolischen Suche durch das AI Team
- Verbessern der Highlights auf Grundlage von Vorschlägen durch das [elastic.co](https://elastic.co) Consulting
- Migration und Verbesserung des GithubStats Enrichments. Erweitern des Enrichment Service für Gitlab
- Kontinuierliche Aufräumarbeiten (für Altlasten) und Ausbau der Dokumentation, da Team schnell wächst.

- Kleine Untersuchung für die Machbarkeit einer "VectorSearch" für Profile → Erfordert besseres Recruiting spezifisches Sprachmodel, welches wir von Keyword Extraction bekommen wollen.