**Stefan x Jan: Update Masterarbeit & Publikation (27.04.2025)**

**Scope-Definition**

* Welche Länder?
  + Statista-Quelle mit den größten Investment-Summen in Europa = United Kingdom, Deutschland, Frankreich, Schweiz, Schweden
  + Aktueller Scope: DACH
  + Zukünftiger Scope: DACH + UK + Frankeich
* Welche Venture Capital Investoren?
  + Investoren mit mindestens 10 Investments
  + Venture Capital Investoren

**Daten-Erhebung**

* Datenquellen
  + Crunchbase (Core-Daten)
  + Pitchbook (Versuch Investorendaten
* Welche Daten werden erhoben?
  + Investorendaten: 486 DACH (989 DACH + UK)
  + Unternehmensdaten: 11,647 Portfolio-Unternehmen
  + Investmentrunden: 40,314 Runden
  + Akquisitionen: 2,875 Akquisitionen
  + Investoren-Offices: Locations/Adressen der DACH-VCs, da über die Hautquellen nur die Hauptanschrift zu ermittlen ist, aber alle weiteren Offices fehlen und dadurch die Distanzberechnungen verfälscht werden würden
  + Founder: Matching mit People Datenset von Crunchbase, um LinkedIn Account zu bekommen und dadurch Profile zu scrapen
* Prinzip Spatial Daten
  + Umwandlung von Adressen in Longitude und Latitude mit der OpenCage API
  + Koordinaten für jedes Unternehmen und jeden Investor
  + Harvesin Formula: Berechnung der Luftdistanz in Kilometer zwischen 2 Koordinaten unter Berücksichtigung der Erdkrümmung

**Daten-Bereinigung & Qualität**

* Entfernen von unnötigen Features, Einführen eines einheitlichen Features-Namens
* Statistische Übersicht über alle Features und Entscheidung, ob ein Feature verwendet werden kann oder nachgebessert werden kann

**Daten-Zusammenführung**

* Zusammenführung von Crunchbase und Pitchbook Daten für Investoren (Datenqualität nicht gut genug)

**Hilfstabellen**

* Investor-Offices
* Investor-Rounds
* Investor-Portfolio

**Daten-Engineering**

* Investoren-Klassifizierung
  + Industrie-Spezialisierung (One-Hot-Encoding aller 50 Industrie-Gruppen, irgendwo mehr als 75%, dann Industrie-Fokus)
  + Spatial-Investortype (regional, dach, europe, worldwide)
* Spatial Features
  + Distanz zwischen Investor und Startup
  + Anzahl/Share Investments im Heimatland
  + Anzahl Investments im Ausland
  + Anzahl/Share Investments in DACH, Europa, Asien, Africa, North-America, South-America, USA, Under-10-KM
  + Unique Countries
  + Durchschnittliche Distanz, Median Distanz
* Exit/Closed-Binaries
* Definitionen von Stages, um Runden zusammenzufassen (z.B. Pre-Seed, Angel, Seed, Equity Crowdfunding, Product Crowdfunding = Seed). Dadurch kann die Entwicklung des Startups gut dargestellt werden, die Transition zu einer weiteren Key-Entwicklungsstufe und die Analyse, ob Stufen übersprungen wurden und wie lange das gedauert hat
* Unclassified Runden werden nach einem Wasserfall-Prinzip anhand der Investmentsumme in die Kategorien einsortiert (z.B. < 5,000,000 = Seed)
* Jahre zwischen den Investmentstufen
* Success-Binaries (IPO, Akquisition) -> perspektivisch auch, wenn Late-stage-Funding
* Anzahl der Investoren-Typen, die ein Unternehmen besitzt, auch nach Funding-Kategorie
* Binaries für die Unternehmen für jede Funding-Kategorie und Investor-Typ
* Follow-on-Investments nachdem ein Investor investiert hat
* Anzahl/Share von Investments der Investoren nach Kategorie
* Herfindahl Index: Hohe Konzentration in einer Investor-Kategorie
* Erfolgsrate je Funding-Kategorie
* Target-VC
* Industry-Novelty
* Alter der Startups
* One-Hot-Encoding der Mitarbeiter-Anzahl
* Distanz zu einem Startup-Hub (Top 30 nach Report)
* Co-Investorenbeziehung: Anzahl-Co-investoren, Anzahl Verbindungen, Tight-Cluster, Bridge-Investor (verbindet Investoren)

**Next Steps**

* Referenzpaper durcharbeiten
* UK und Frankreich Daten erheben
* Daten für Regressionen vorbereiten
* Modellannahmen treffen, belegen und Modell visualisieren
* Regressionsanalyse durchführen (verschiedene Settings, Multilevel, logistische Regression)