Anforderungskatalog zur Definition des Datascience Life Cycles

Github Test

# Definition des Daten Managements (Datenformate, Datenstruktur, Datendichte, Datenablage, verwendete IT-Tools und deren Versionierung, Zugriffsschema)

Ideen

-PosgresSQL wenn möglich

-Schnell „Spiel“ Daten Sammeln

-Server mit Zugriff aufsetzen (Hetzner?)

-Elegantere langfristige Lösung recherchieren.

Fragen

-Welche Messwerte werden erhoben, in welchen abständen, wie werden Daten gespeichert/Weitergereicht

ToDos:

-Rauskriegen wie welche Daten gemessen werden sollen

-Wie können daten gespeichert und übergeben werden

-Dann Technologie recherchieren vorher erstmal mit bekanntem arbeiten.

# Definition der Git/GitHub und One-Drive-Struktur als zentrale Datenaustauscheinheiten

Ideen

-Entwerder richtig gitlabs auf server oder normal Github

Fragen

-Wer Produziert alles code für Github?

-Besondere Anforderungen an Datensicherheit oder Funktionalität

ToDos

* Mit Igor besprechen
* Github Account aufräumen und einarbeiten

# Definition von ETL-Pipeline (Data Flow Diagramm, Datensemantik, Transformationsschema)

Ideen

Fragen

-Zuerst punkt 1 klären

-Was soll Lokal passieren was im Internet?

-Begrenzung kapazitäten bei Lokalen Controllern/Rechnern

ToDos

-Technologie Recherchieren

-Anforderungen und Limitierungen mit anderen absprechen

# -Definition des Scrapping-Verfahrens, Labelingschema der Daten)

Ideen

-Sollte relativ unabhängig realisierbar sein bei klaren Vorgaben

Fragen

-Welche Daten sollen gescraped werden?

-Was muss gelabelt werden (Stichwort Zielgrößen)?

-Wie häufig soll aktualisiert werden

ToDos

-Rauskriegen welche Daten

-Einmaliger Download und sichten der Daten

-Wenn Sinnvoll Automatisierung und Einbindung in DB

# Definition der Cloud-Struktur (Relationen der Datenbank,

Ideen

-Einfache Lösung: Server oder VM je Kunde mit eigener DB+Vortrainiertes Modell (An Labor und Feldversuch Daten)

-Komplexe Lösung: Ein skalierbares, Globales System

Fragen

-Zuerst 1 und 3 klären.

-Wie soll das Endprodukt aussehen? SaaS (schwierig) oder separate Serverinstallation für jeden Kunden (einfacher/sicherer aber nicht so richtig Cloud).

ToDos

-Technologie recherchieren für einfache und komplizierte Lösung

-Grundsätzliches Schema erarbeiten

# Definition der stat. Kennzahlen in der Versuchsplanung (Art des Versuchsplanes, Definition der Variablen, Standardabweichungen, Fehlerfortpflanzungen, stat. Power, FDS, Korrelationsgüte, p-Wert, AIC, BIC)

Ideen

-Standard Bewertungskriterien für Versuchspläne, eventuell zusätzlich Simulations Designs

-Zusätzlich Leitfaden für Projektablauf definieren(Anforderungen an Faktoren Zielgrößen Messsysteme etc)

-Einstellung des Systems, eventuell Fehler provozieren/vorgeben (vgl. salmaso2019-DoEPredMaintenance)

Fragen

-Wie sehen die Laborumgebungen aus Teilsysteme Gesamtsystem beides

-wird es auch Computer Simulationen geben.

-Zielgrößen (siehe 1)

ToDos

-Projektleitfaden

-Bewertung von DoEs und Modellen

-Fragen klären

# Definition der Qualitätskriterien für die KI-Algorithmen (Stichprobenumfang, Güte der Voraussage, Güte des Models, Geschwindigkeit der Code-Ausführung)

Ideen

-Standardbewertungskriterien

Fragen

-Zielgrößen?!

-Welche Verfahren? Nur Supervised?

-Zeitreihen- Supervised; Trenderkennung strukturbrüche etc

ToDos

-Ziele und Zielgrößen klären

-Recherche für Zeitreihen sowohl Modellierung/bewertung im Supervised bereich als auch Unsupervised (Trends etc.)

# Anforderungen an Visualisierung der Daten (Ausführungszeit vs. Datenmenge, eindimensional vs. verschachtelt)

Ideen

-Zeitreihen von bestimmten Sensordaten (z.B. 1h).

-Modelgrafiken

-Wissenschaftlich Standards einhalten (z.B. Grammar of graphics)

Fragen

-Visualisierung auch im Monitoring dashboard (eher nicht, nur KPIs)?

ToDos

-Beispiel UI suchen und Funktionalität sammeln.

1. Definition der Automatisierung, z.B. zentrale Bündelung der Daten auf IntelNuc/RapsberryPi (Rechnergeschwindigkeit, Speicherplatz etc.), Datenbank (Berechtigung, relationale Schema), Automatisierung des Versuchs z.B. Python/Labview (Versuchszeit, Ausführungsgeschwindigkeit, Datenablage)

* **Vgl.1. und 3.**
* Wann fängt unser zuständigkeitsbereich an?
* In welchem Format können die Daten geliefert werden, in wie fern können verschiedene Messwerte einer Messung zugeordnet werden (nur Zeitstempel oder sind Sensoren gleichgeschaltet?)
* DB Organisation für Beispieldaten am Anfang, und perspektivisch.