# Bachelorarbeit/Masterarbeit: Entwicklung und Anwendung von Data Science Techniken für ungelöste Informatik-Probleme

Es gibt eine faszinierende Liste offener Mathematik- und Informatik-Probleme, s. b. https://de.wikipedia.org/wiki/Ungel%C3%B6ste_Probleme_der_Mathematik oder https://de.wikipedia.org/wiki/Liste_ungel%C3%B6ster_Probleme_der_Informatik. Dazu gehören z. B. Fragen nach der Komplexität bestimmter Fragestellungen oder des Entwurfs effizienter Algorithmen. In dieser Arbeit geht es darum herauszufinden, inwieweit Big Data Methoden helfen können, ein besseres Verständnis von solchen Problemen zu bekommen (nicht in erster Linie darum, sie zu lösen!), und wie entsprechende Systemunterstützung aussehen könnte. Dies anhand eines Beispiels (das nicht auf einer der o. g. Listen steht, aber vergleichsweise anschauliche und kurze Erklärungen erlaubt) in etwa wie folgt:

Sogenannte kinetische Datenstrukturen, z. B. kinetic heaps, s. b. https://en.wikipedia.org/wiki/Kinetic_heap, verwalten Werte, die sich über die Zeit ändern. D. h. die Antwort beispielsweise auf die Frage, welche Variable den kleinsten Wert enthält, hängt vom Zeitpunkt ab. Das Element, das in einem Heap derzeit ganz oben steht, wird also i. Allg. irgendwann in der Zukunft durch ein anderes ersetzt. Eine sogenannte Event Queue verwaltet die Zeitpunkte, zu denen die kinetische Datenstruktur derart reorganisiert werden muss. – Im vorliegenden Kontext ist jetzt wichtig, dass in bestimmten (überraschend einfachen) Fällen die Komplexität des kinetic heaps nicht bekannt ist.

In einem ersten Bearbeitungsschritt implementieren Sie zunächst einen kinetic heap. Sie generieren dann automatisch sehr viele Beispiele. Ein Beispiel ist hier eine Menge sich über die Zeit ändernder Werte. Sie erfassen dann ‚breit und ausführlich' für jedes Beispiel die Kosten, die anfallen, wenn ein kinetic heap diese Daten verwaltet. ‚breit und ausführlich' bedeutet hier, dass Sie beispielsweise nicht nur einfach die Gesamtanzahl ausgeführter Operationen erheben, sondern fein aufgeschlüsselt z. B. abhängig von der Tiefe der ursprünglichen Position in der Event Queue, von der Anzahl der Operationen, die unmittelbar vorangegangene Events verursacht haben usw. Diese Kennzahlen/unterschiedlichen Kosten jedes Beispiels dienen dann als Eingabe für die sich anschließende Datenanalyse; sie werden üblicherweise als Merkmale bezeichnet. Darüber hinaus gibt es noch Merkmale, die die Struktur des Beispiels beschreiben, hier z. B. die maximale Geschwindigkeit, mit der sich ein Wert eines Beispiels ändert. Sie sollen sich darüber hinaus aber auch selbst möglicherweise aufschlussreiche Merkmale überlegen. Auf all diese Merkmalsdaten wenden Sie dann existierende Verfahren für die Datenanalyse an. Das tun Sie, um u. a. folgende Fragen zu beantworten:

- Was ist an den Beispielen, für die die Gesamtkosten am größten sind, besonders? Haben sie irgendwelche strukturellen Gemeinsamkeiten?
- Welche Merkmale haben eine besonders ausgeprägte Korrelation? Welche Kombinationen von Merkmalswerten, die auf den ersten Blick plausibel/möglich zu sein scheinen, kommen nie vor?
- Gibt es in der Menge der Beispiele Outlier und Cluster? Wenn ja, was bedeuten sie anschaulich?

Sie sollen sich noch ein, zwei weitere Probleme (d. h. etwas anderes als kinetic heaps) vornehmen und Merkmale, die die strukturelle Seite des Problems beschreiben, als auch Merkmale, die die jeweilige ‚Lösung' (im Kontext oben die Verwaltung konkreter Daten mit einem kinetic heap) charakterisieren,

überlegen, das Ganze implementieren, jeweils viele Daten systematisch erheben, sie strukturiert ablegen und analysieren.

Eine Besonderheit dieses Anwendungsfalls aus Data-Science Sicht sehen wir darin, dass man grundsätzlich ohne Limitierung neue Daten generieren und/oder neue Merkmale hinzunehmen kann. Zu den Einsichten, die wir uns von Ihrer Arbeit erhoffen, gehören deshalb die folgenden:

- Wie sollte Systemunterstützung für diese Art von Data Science-Prozess aussehen? – Wenn Sie Bachelorarbeiter sind, müssen Sie ihm Rahmen Ihrer Bearbeitung kein derartiges System bauen, dann reichen uns die Einsichten; von einem Masterarbeiter wünschen wir uns hingegen eine zumindest rudimentäre prototypische Implementierung. Diese Differenzierung gilt auch für das folgende Item.
- Welche Interaktionsmöglichkeiten mit einem entsprechenden System sollte ein Anwender haben? (D. h. was für eine Benutzerschnittstelle ist sinnvoll? Welche Arten von Feedback sollte ein Anwender zu den Analyseergebnissen geben können?)
- Wie sieht die Rückkopplung aus? D. h. welche Anforderungen an Datenanalyseverfahren ergeben sich aus den möglichen Arten von Benutzerfeedback? (Hier reichen uns in jedem Fall die Einsichten, sei es eine Bachelor-, sei es eine Masterarbeit.)

D. h. Beurteilungskriterien für Ihre Arbeit sind:

- Wie sauber gehen Sie bei der Ausgestaltung des oben beschriebenen Ablaufs und der Implementierung jeweils vor?
- Wie findig sind Sie, was Vorschläge für Merkmale angeht? (Dieses Kriterium ist allerdings von untergeordneter Bedeutung – wir haben hierzu bereits einige Vorschläge.)
- Wie umfangreich und tiefgehend sind Ihre Antworten auf die o. g. Fragen, die Einsichten betreffend?

Diese Aufgabenstellung ist also im Kern eine Data Science Aufgabenstellung, allerdings mit einem unseres Wissens weitgehend unerforschten Anwendungsfall.

Betreuer: Klemens Böhm, klemens.boehm@kit.edu