

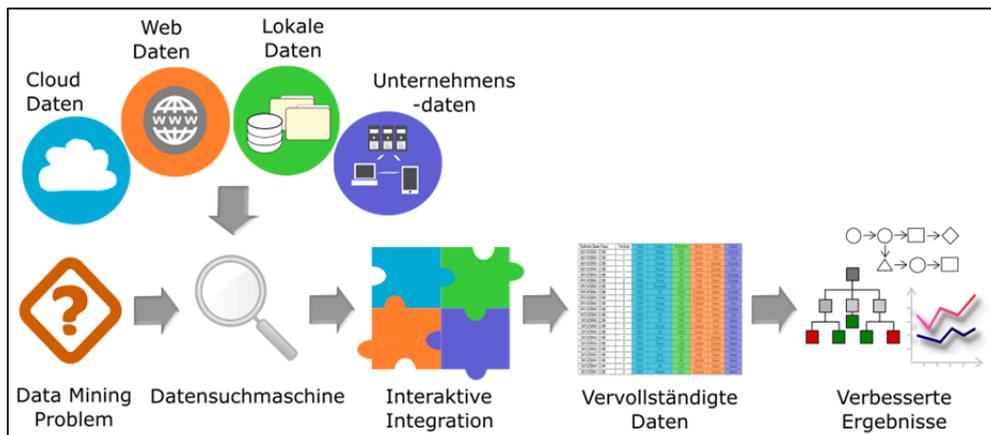
# Datensuche für Data-Mining-Prozesse (DS4DM)

**Projektpartner: Rapidminer GmbH und Universität Mannheim**

**Problemstellung:** Die Menge der verfügbaren Daten innerhalb von Unternehmen sowie im öffentlichen Web ist in den letzten Jahren sprunghaft gestiegen. Aufgrund der sehr großen Anzahl an Datenquellen sehen sich Analysten zunehmend mit der für sie sehr unbefriedigenden Situation konfrontiert, dass die Daten, die sie für ihr Data-Mining-Projekt benötigen, zwar irgendwo innerhalb des Unternehmens oder im Web vorliegen, sie aber nur schwer auffindbar sind und mit großem Aufwand integriert werden müssen. Der Fokus vieler Data-Mining-Projekte verschiebt sich somit zunehmend von der eigentlichen Datenanalyse auf die Suche nach geeigneten Daten zur Beantwortung der jeweiligen Fragestellung sowie deren Integration.

**Lösungsansatz:** Die derzeit am Markt angebotenen Data-Mining-Werkzeuge unterstützen Analysten kaum bei der Suche nach geeigneten Datensätzen sowie bei der Integration relevanter Daten aus verschiedenen Datenquellen. Im Projekt DS4DM soll RapidMiner, eine der weltweit führenden Datenanalyse-Umgebungen, um Funktionalität zur Suche und Integration von Daten aus unternehmensinternen Datenquellen sowie aus dem öffentlichen Web erweitert werden. Mit Hilfe der neuen Datensuch- und Datenintegrations-Funktionalitäten lassen sich die Kosten und die Dauer von Data-Mining-Projekten deutlich reduzieren.

**Kernarbeiten des Projektes:** 1. Entwicklung von Methoden für die korrelations-basierte Datensuche sowie Bereitstellung dieser Methoden in Form einer Datensuchmaschine; 2. Entwicklung einer interaktiven Datenintegrationsumgebung, mittels der sich lokale Tabellen um zusätzliche Attribute basierend auf den gefundenen Datensätzen erweitern lassen; 3. Entwicklung von Adaptern, die es der Datensuchmaschine ermöglichen, auf unternehmensinterne sowie Web-Datenquellen zuzugreifen und diese für die Suche zu indexieren.



**Verbesserung von Data Mining Prozessen mittels der Suche und Integration zusätzlicher Daten**

**Verwertungspotential:** Das Projekt positioniert die Rapidminer GmbH als ersten deutschen Softwareanbieter im Innovationsfeld der Datensuche, in dem bisher nur amerikanische Firmen wie Google und Microsoft aktiv sind. Die im Projekt prototypisch entwickelte Funktionalität wird nach Projektende als RapidMiner Server Modul sowie als Software-as-a-Service-Lösung vermarktet werden. Durch die Veröffentlichung der Methoden in Form von wissenschaftlichen Publikationen sowie die Bereitstellung der Datensuchmaschine mittels eines OEM-Modells können die Methoden auch von weiteren Softwareanbietern im Bereich Data Mining sowie Enterprise-Content-Management genutzt werden. Die entwickelten Methoden ermöglichen es Unternehmen, unternehmensinterne Daten sowie Web-Daten umfassender für die Entscheidungsfindung zu nutzen.