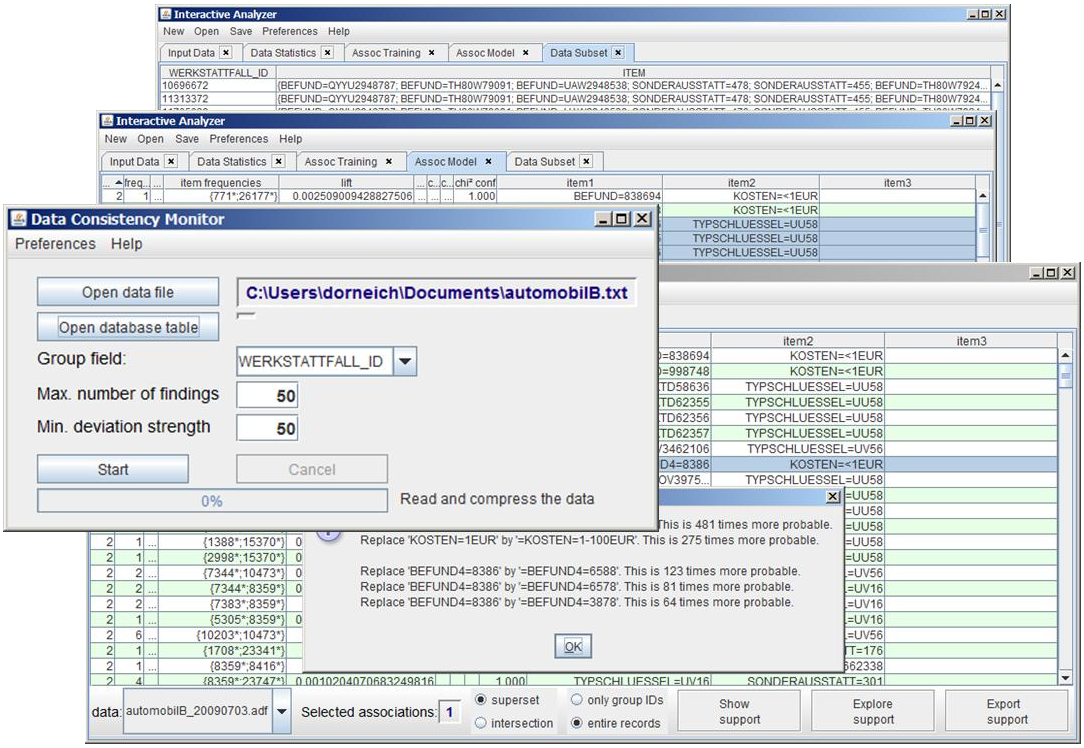
# interactive analyzer

understanding data

## Proaktives Datenqualitätsmanagement

Basierend auf fortschrittlicher Data-Mining-Technologie findet der Synesis Interactive Analyzer völlig autonom und ohne manuelle Definition von Regeln, Filtern oder Kriterien selbst hoch komplexe Fehlermuster in Ihren Daten. Die Software liefert Korrekturvorschläge und unterstützt Sie bei der raschen Behebung der Inkonsistenzen. Einfache Bedienung, vielfältige Integrationsmöglichkeiten und das flexible Lizenzmodell machen die Software für Unternehmen jeder Größe attraktiv.




### Business Case

Die permanente Sicherstellung von Datenqualität und Datenkonsistenz geschäftsrelevanter Datensammlungen ist eine grundlegende Herausforderung für Manager und Fachkräfte aus IT, Qualitätssicherung, Logistik, Kundenmanagement und vielen anderen Bereichen. Experten schätzen die von fehlerhaften Daten verursachten volkswirtschaftlichen Gesamtkosten alleine in den USA auf mehr als 600 Milliarden US $ jährlich (Eckerson, 2002): Datenfehler können fehlerhafte Produkte und Dienstleistungen hervorbringen, die Effektivität von Marketing- und Vertriebskampagnen verringern oder die Ursache für mangelhaftes Kundenmanagement und unzufriedene Kunden sein.

Aus diesem Grund wurden vielerorts unternehmensweite oder bereichsspezifische ‘Data Governance’ Funktionen geschaffen, deren Hauptaufgabe die permanente Einhaltung festgelegter Datenqualitäts-Standards ist. Dennoch besteht weithin eine gewaltige Diskrepanz zwischen den hohen Geldsummen, die für Hardware, Software, Administration und Datenerfassung ausgegeben werden, und der Aufmerksamkeit, die der Qualität der erfassten Daten gewidmet wird. In Summe führt das zu ineffizientem Mitteleinsatz und Unzufriedenheit mit den Systemen.

Datenqualitätsprobleme entstehen durch unvollständige, unrichtige oder inkonsistente Daten. Eine Software-basierte Lösung zum Datenqualitäts-Monitoring sollte daher einige oder alle der folgenden Funktionen beinhalten:

- **‚Data Profiling'** – Datenüberprüfung mit dem Ziel, Probleme und Herausforderungen zu identifizieren.
- **Daten-Standardisierung** – ein Regelsystem für Datenaufnahme und –verarbeitung, das sicherstellt, dass die Daten definierte Qualitätsregeln einhalten.
- '**Matching or Linking'** – systematische Datenabgleiche, so dass ähnliche aber leicht unterschiedliche Datensätze zusammengeführt werden können (**Dublettenbereinigung**).
- **Monitoring** – systematische und regelmäßig ablaufende Datenqualitätsüberwachung und Generierung von Berichten oder (semi-)**automatisch ablaufende Korrekturmaßnahmen**

[Dieser Text basiert auf dem Wikipedia-Artikel http://en.wikipedia.org/wiki/Data_quality]

## Datenqualitäts-Monitoring mit dem Interactive Analyzer

Traditionelle Methoden zur Aufdeckung und Behebung von Datenqualitätsproblemen betrachten einzelne Datenmerkmale auf fehlende oder aus dem üblichen Wertebereich herausfallende Werte. Ferner werden manuell definierte Filter und 'Business Rules' eingesetzt. All diese Methoden unterstützt der Interactive Analyzer auch. Als Alleinstellungsmerkmal werden darüber hinaus Data-Mining-basierte Methoden angeboten. Diese finden völlig selbständig, ohne dass manuelle Filter und Business Rules definiert werden müssen, subtile Inkonsistenzen zwischen den Werten verschiedener Merkmale eines Datensatzes. Zu jeder gefundenen mutmaßlichen Inkonsistenz findet die Software schließlich die – statistisch begründet – besten Korrekturvorschläge. Die Vorteile des Interactive Analyzer Ansatzes sind:

- **Automatische Fehlersuche ohne Voreinstellung von Fehlerkriterien:** Die Software findet eigenständig die Inkonsistenzkriterien, sie findet **alle statistisch** in der Datenbasis **signifikanten Kriterien** und sie findet **nur die statistisch signifikanten Kriterien**. Dieser 'Hypothesen-freie' Ansatz erspart nicht nur viel Arbeitszeit, er stellt auch sicher, dass kein wichtiges Kriterium bei der Definition der anzuwendenden Filter-Kriterien übersehen wird, dass aber auch nicht ‚nach Bauchgefühl' Inkonsistenzkriterien definiert werden, die sich auf der Datenbasis gar nicht rechtfertigen lassen.
- **Automatische Identifikation von ‚Top-Problemen':** Alle gefundenen mutmaßlichen Inkonsistenzen werden mit objektiven statistischen Signifikanzzahlen bewertet und quantifiziert, so dass sich geordnete Listen von 'Top-Problemen' bilden lassen. Dadurch wird die Begründung der Einstufung als Inkonsistenz und der vorgenommenen Korrektur reproduzierbar, auditierbar und jederzeit nachvollziehbar und kann automatisch protokolliert werden – zum Beispiel um regulatorische Vorgaben zu erfüllen.
- **Identifikation hochkomplexer Datenzusammenhänge:** Die eingesetzten Data-Mining-Verfahren können auch Muster und Zusammenhänge entdecken und verwenden, die wegen Ihrer Komplexität oder wegen der Größe oder Komplexität der Datenbasis mit traditionellen Methoden nicht entdeckt werden können.

Insgesamt ist der Data-Mining-Ansatz des Interactive Analyzer sehr gut komplementär zu bestehenden Methoden der Datenqualitätssicherung einsetzbar: er füllt die Lücken, die andere Methoden auf den Gebieten der Objektivität, Vollständigkeit, Analysegeschwindigkeit, Nachvollziehbarkeit, Dokumentation und Automatisierbarkeit haben.

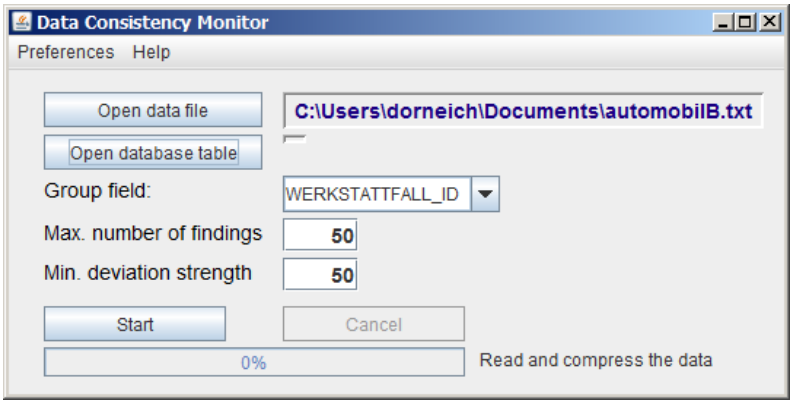
## Alleinstellungsmerkmale des Interactive Analyzer

Im Vergleich zu anderen Methoden und Software-Werkzeugen aus dem Bereich 'Data Quality Assurance' bietet Interactive Analyzer die folgenden Differenzierungsmerkmale:

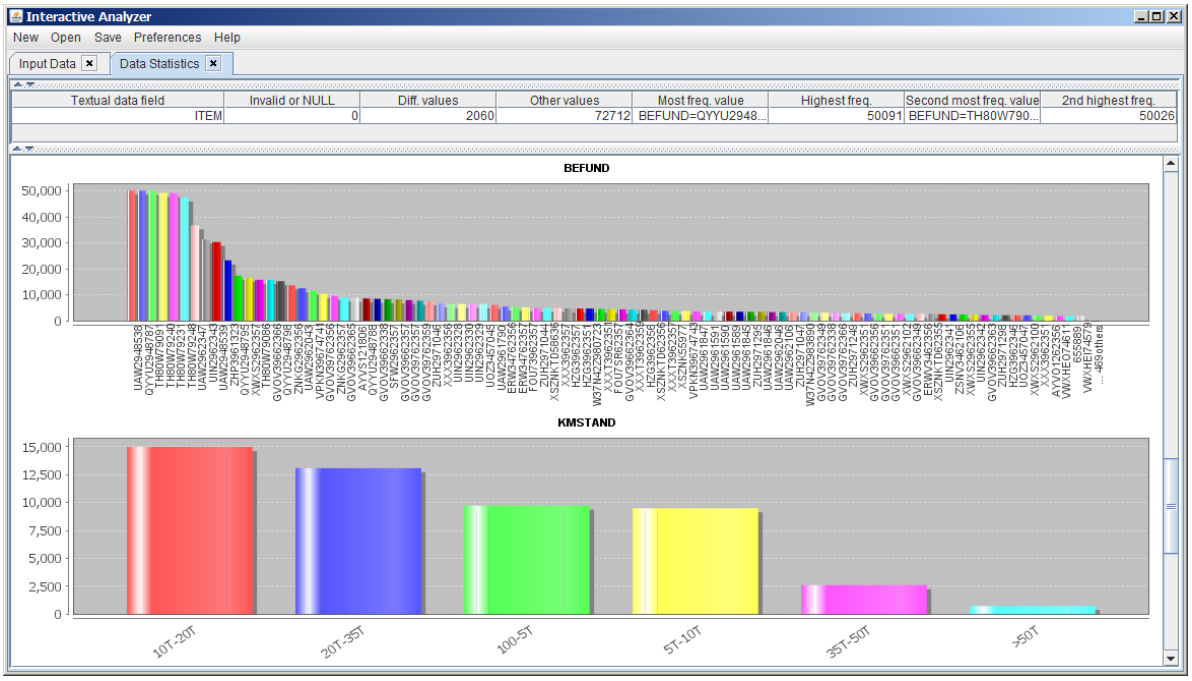
- Völlig **autonom**, ohne manuelle Definition von Regeln oder Kriterien, findet Interactive Analyzer auch **hoch komplexe Fehlerzusammenhänge**.
- **Leichte Bedienbarkeit**: Einfache Oberflächen, leichter Zugriff auf Datenbanktabellen, Datenbank-Sichten, Excel-Sheets oder Flachtextdateien. **Schneller Überblick über die Güte der Datenqualität**
- **Sehr günstige Total Cost of Ownership**: die geringen Hardware-, Administrations- und Einarbeitungsanforderungen und das flexible Preismodell machen den Einsatz des Interactive Analyzer für Unternehmen jeder Größe interessant.
- **Geschwindigkeit und Interaktivität**: die Datenkonsistenzanalyse dauert typischerweise nicht länger als einige Minuten.
- **Leistungsfähigkeit und Skalierbarkeit** durch seine komprimierte 'In-Memory'-Datenhaltung und seine leistungsfähigen, parallelisiert ablaufenden Analyse-Algorithmen.
- **Integration in bestehende IT-Strukturen und Prozesse**: Interactive Analyzer interagiert mit Datenbanken, Reporting-Systemen und Unternehmensapplikationen über standardisierte Schnittstellen wie JDBC, Web Services (SOA) und XML. Die Software ist nutzbar in Form von
    - automatisierten Prozessen und regelmäßig ablaufenden Batch-Job, Systemservices oder Datenbankprozessen.
    - selbständigen Anwendungsprogramme mit intuitiven grafischen Benutzeroberflächen
    - Java Applets, in Unternehmensportale und Reporting-Anwendungen integriert
    - Web Services im Rahmen von SOA-Architekturen
    - Funktionalen Erweiterungen ('stored procedures') in Oracle, IBM DB2 usw.

### Anwendungsbeispiel: Garantie- und Kulanz-Daten eines Automobilherstellers

Im folgenden Anwendungsbeispiel wird eine Tabelle mit Hilfe des Interactive Analyzer ™ Data Consistency Monitor auf Datenqualität, Datenkonsistenz und Auffälligkeiten untersucht, die Daten zu Werkstattaufenthalten von Fahrzeugen im Rahmen von Garantiefällen oder Kulanzreparaturen enthält. Zu jedem Werkstattfall enthält die Tabelle mehrere Datenzeilen mit Informationen über das Fahrzeug (Typschlüssel, Sonderausstattungen, km-Stand usw.), den internen Fehlerspeicher-Status zum Reparaturzeitpunkt sowie die abgerechneten Reparaturbefunde und –kosten. Die Daten sind OEM unabhängig und künstlich erzeugt, wurden aber realen Anwendungsprojekten in der Automobilindustrie nachempfunden.

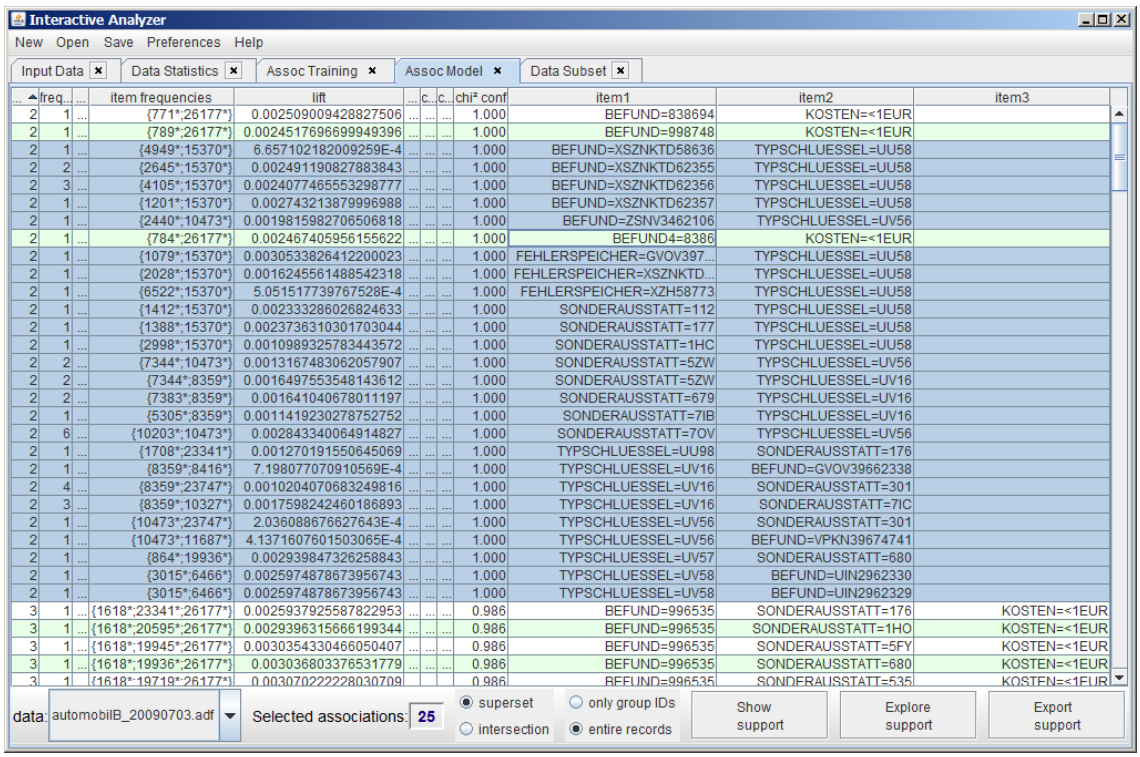



1. In einem ersten Schritt verschaffen wir uns einen Überblick über die in den Daten enthaltenen Informationen und die statistischen Verteilungen der einzelnen Merkmalswerte. Die Daten erscheinen gut gepflegt: es sind weder größere Datenlücken noch fehlerhafte Merkmalswerte erkennbar.

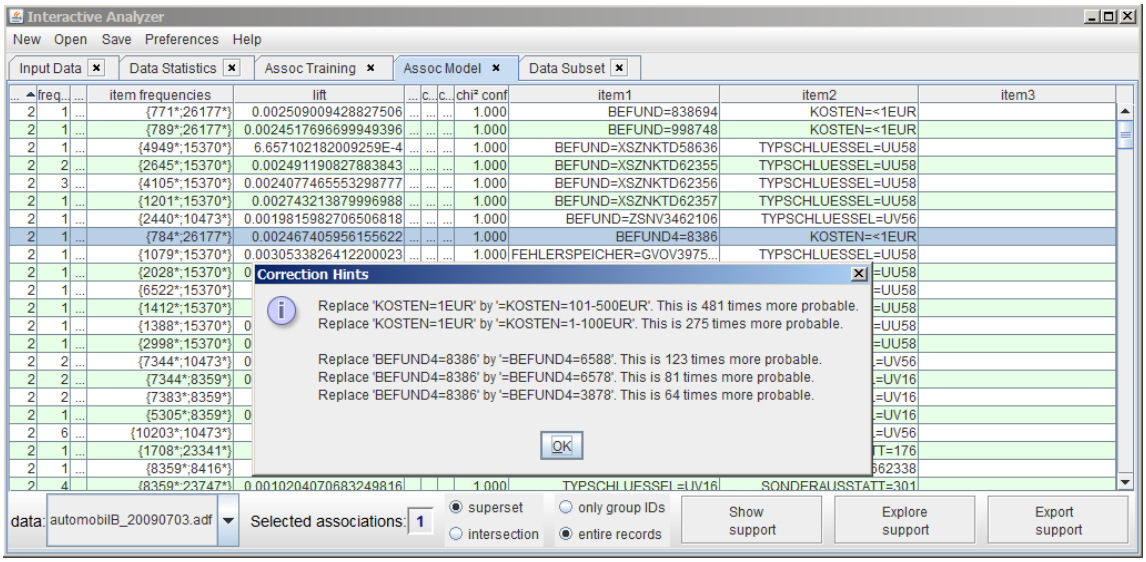



2. Nun wollen wir mit Hilfe des Data-Mining-Ansatzes herausfinden, ob dennoch Inkonsistenzen in den Daten enthalten sind. Tatsächlich findet Interactive Analyzer Dutzende von Auffälligkeiten (siehe folgende Abbildung). Die ersten ca. 25 Auffälligkeiten wurden allesamt als hoch-signifikant eingestuft (der chi²-Konfidenztest liefert eine Verlässlichkeit von >99.9%). Eine Teilmenge dieser Auffälligkeiten betrifft Kombinationen von Fahrzeug-Baureihen (Typschlüsseln), Befund-Schlüsseln und Sonderausstattungen, die es in dieser Kombination eigentlich nicht geben dürfte. Diese Muster wurden durch Anklicken mit der Maus blau markiert.

3. Mit Hilfe der drei Knöpfe ‚Show Support', Explore Support' und ‚Export Support' lassen sich die betroffenen Fahrzeuge und Reparaturvorgänge in den Originaldaten anschauen, in Charts visualisieren sowie in eine Textdatei, ein Excel-Sheet oder eine Datenbank exportieren, um die Daten gegebenenfalls zu korrigieren.



Diese Korrektur kann manuell, das heißt durch einen menschlichen Fachexperten, erfolgen. Interactive Analyzer kann aber auch eigenständig, aufbauend auf einer statistischen Assoziationsanalyse der Datenbasis, Korrekturempfehlungen vorschlagen. In der folgenden Abbildung werden Korrekturempfehlungen für die blau markierte Auffälligkeit angezeigt. Offensichtlich wurde in einem Reparaturfall ein Reparaturbefund (‚8386'), der normalerweise mit Kosten von knapp unter 100 € (275 mal) oder knapp über 100 € (481 mal) verbucht wird, fälschlicherweise als kostenfrei verbucht (Kosten<1€).




## Zusammenfassung

Interactive Analyzer findet Auffälligkeiten, Datenfehler und Inkonsistenzen, die sich erst in der Zusammenschau mehrerer Datenmerkmale als Auffälligkeit herausstellen und deshalb von den gängigen Datenqualitäts-Überwachungsmethoden nicht gefunden werden. Darüber hinaus liefert Interactive Analyzer durch quantitative Analysen begründete Korrekturhinweise und erlaubt so eine effektive, automatisierte oder zumindest Software-unterstützte Behebung von Datenfehlern. Und die ersten aussagefähigen Ergebnisse erhalten sie meist schon am ersten Einsatztag, ohne komplexe Hardwareanforderungen und überbordende Lizenzkosten.

## Kontakt

Dr. Ansgar Dorneich
Fichtenstraße 7
71088 Holzgerlingen
Email: info@i-analyzer.com
Web: www.i-analyzer.com