# Matthias

## Einleitung

**Folie - Titelfolie**

Wirtschaft im Beziehungscheck. Das ist der Titel des Bachelorprojektes, an dem unsere Gruppe seit 10 Monaten arbeitet.

Wirtschaft im Beziehungscheck heißt, wir wollen mehr erfahren über deutsche Unternehmen und wir wollen mehr erfahren über die Beziehungen, die zwischen Unternehmen bestehen.

Und warum interessiert uns das? Risikomanagement für die Commerzbank.

Kurz definiert: Risiko ist die Möglichkeit Wert zu gewinnen oder zu verlieren.
Und Risikomanagement ist die Steuerung dieser Risiken. Also beispielsweise das Kreditausfallrisiko zu mindern.

Was haben Unternehmens-Beziehungen damit zu tun? Beispiel: Wir schreiben das Jahr 2008. Finanzkrise/Weltwirtschaftskrise. Eine häufig gehörte Frage war, welche Unternehmen werden durch die Lehman Brothers Pleite noch mit in den Abgrund gezogen. Es geht also um versteckte Risiken.

Nun bei großen bekannten Unternehmen mögen diese Zusammenhänge offenkundig sein, aber bei über 4 Millionen Unternehmen in Deutschland verliert man schnell den Überblick.

Wie kann man nun Übersicht in die Beziehungsgeflechte bringen. Lasst uns vorstellen wir sind Kundenberater bei der Commerzbank ein fiktives Beispiel nehmen:

- Die “Bunte Monokulturen im Raubbau AG” beliefert die “Katzis klebrige Kauklumpen GmbH & Co KG” mit lustigen Farben für die Kaugummiproduktion.
- Die “Quick’n’Dirty Zeitarbeit AG” ist Kundin der Commerzbank AG und hat zur Zeit alle ihre 10.000 Arbeitsminions an “Katzis klebrige Kauklumpen GmbH & Co KG” verliehen.
-

Jonathan: Doch was passiert jetzt?! [Ereigniskarte überreichen]
Matthias: Au weia!

**Folie - Monopoly Ereigniskarte**
[Animation: Ereigniskarte aufdecken]

“Im Anbaugebiet für lustige Farben regnet es seit Wochen. Der Himmel ist ständig grau. Die ‘Lustige Farben’-Ernte ist im Eimer. ‘Katzis klebrige Kauklumpen GmbH & Co KG‘ gehen die

Zutaten aus und kann den Kredit nicht an die Commerzbank AG zurückzahlen. Alle Arbeitsminions der Quick'n'Dirty Zeitarbeit AG werden sofort freigestellt und offene Rechnungen für bereits geleistete Minionarbeit bleiben offen. Gehen Sie mit Ihrem Kunden zum Insolvenzverwalter. Gehen Sie dabei nicht über 'Los' und ziehen Sie nicht 4.000€ ein."

# Jonathan

Wie konnte das passieren? Ihnen ist entgangen, wie sehr die "Quick'n'Dirty Zeitarbeit AG" von der "Lustige Farben"-Ernte abhängig ist. Was Ihnen gefehlt hat, war eine Übersicht der Beziehungen zwischen den Unternehmen - ein Beziehungscheck für die Wirtschaft.

**Folie - Konzept eines Netzwerkgraphen**
Eine geeignete Darstellung der Unternehmensbeziehungen ist dieser Netzwerkgraph [zeigen]. Er besteht aus den Unternehmen - hier als Logos dargestellt - [zeigen] sowie ihren Beziehungen zueinander – hier als Pfeile dargestellt [zeigen]. Übertragen wir diese Art der Darstellung auf alle deutschen Unternehmen [Einblendung], so entsteht ein enorm umfangreiches Netzwerk.

**Folie - Unternehmensnetzwerk**
Es besteht aus 4 Millionen Unternehmen. Hier erhalten Sie einmal einen visuellen Eindruck davon, von welcher Größe und Komplexität dieses Geflecht ist. Hinzu kommt, dass wir es mit einer dynamischen Unternehmenslandschaft zu tun haben, die sich fortlaufend verändert.
Wir haben uns zum Ziel gesetzt, das Netzwerk der deutschen Unternehmen für die Risikomanager der Commerzbank in Echtzeit zu ermitteln und übersichtlich darzustellen.

# Lösung

**Folie - Software als Maschine**

Dazu haben wir in den letzten 10 Monaten eine Software entwickelt, die aus öffentlich verfügbaren Quellen [Einblendung] den Netzwerkgraphen aufbaut und auf dem aktuellen Stand hält [Animation].

**Folie - Quellen**

Wir verwenden zwei Arten von Quellen: Strukturierte und unstrukturierte.
Strukturierte Quellen enthalten maschinenverständliche Tabellen. Wir beziehen sie aus [Einblendung] Wikidata, DBpedia, Kompass und Implisense.
Unstrukturierte Quellen enthalten Fließtexte, was in unserem Fall deutschsprachige Artikel aus [Einblendung] beispielsweise Wikipedia oder Spiegel Online sind.

**Folie - Software als Maschine (geschlossen)**

Was macht unser System nun mit diesen Quellen? Nun, wir haben es in drei Komponenten aufgeteilt:

**Folie - Blick in die Maschine (Agenda)**

Die erste Komponente [Einblendung], die Data Integration, verarbeitet die strukturierten Quellen.
Die zweite Komponente [Einblendung], die Data Extraction, verarbeitet - Richtig geraten! - die unstrukturierten Quellen.
Die dritte Komponente [Einblendung], das Data Cockpit, dient zur Steuerung und Überwachung der Software.
Diese drei Komponenten wollen wir Ihnen jetzt im Detail vorstellen.

# Matthias

## Data Integration

**Folie – Blick in die Maschine mit Blick in die Data Integration**
Die erste Komponente, die Data Integration, nimmt strukturierte Daten entgegen und führt damit die folgenden drei Schritte durch:

1. Normalisierung
2. Entfernung doppelter Einträge
3. Konstruktion des Netzwerkgraphen

**Folie – 1. Schritt Normalisierung**

Die strukturierten Unternehmensdaten kommen aus verschiedenen Quellen. Sie sind daher uneinheitlich formatiert. Das wollen wir ändern:

Hier sehen wir beispielhaft verschiedenen Schreibweisen des Umsatzes des Volkswagen-Konzerns. Zur besseren Vergleichbarkeit transformieren wir diese Daten in ein einheitliches Format.

*217 Mrd. Umsatz*
*217.267.000.000 EUR*
*2.172 euro gross profit on sales*
*200 billion € turnover*
*zweihundert Milliarden Euro Absatz*

**Folie - 2. Schritt: Doppelte Einträge entfernen**
Als nächstes sucht diese Komponente nach Unternehmen, die doppelt oder sogar vielfach in unseren Daten auftauchen.

Wir haben inzwischen rund 2 Millionen Unternehmen in unserer Datenbank. Um alle Dopplungen zu finden, müsste die Software einem naiven Ansatz nach alle Unternehmen paarweise miteinander vergleichen. Das ergäbe Billionen von Unternehmensvergleichen überall deren verschiedene Eigenschaften.

Bei einer durchschnittlichen Rechenperformance würde das Auffinden von Dopplungen auf diese Weise rund 35 Jahre dauern. Das entspricht 306.000 Stunden.  Solange wird kein Risikomanager auf ein Ergebnis warten wollen, und nebenbei wäre es bis dahin wohl auch veraltet.

Deswegen haben wir ein heuristisches Verfahren implementiert, dass unsere Unternehmensdatenbank clever in sehr viel kleinere Teildatensätze zerlegt. Vergleiche werden nur noch innerhalb dieser Teilmengen durchgeführt. Ergebnis: Es dauert nur noch 1h. [Animation 35 Jahre auf 1h runterzählen lassen]

Wir gewinnen mit unserer Software dadurch zügig einen *fast* dopplungsfreien Netzwerkgraphen der deutschen Unternehmenslandschaft.

**Folie - Netzwerkgraphgenerierung**
Die Quellen enthalten neben den Informationen zu Unternehmen auch Angaben zu ihren Beziehungen zu anderen Unternehmen. Daraus baut unsere Software den Netzwerkgraphen zusammen.

**Folie – Blick in die Maschine (Agenda)**
Nun hat unsere Software ein Unternehmensnetzwerk aus den strukturierten Quellen aufgebaut. Was passiert eigentlich mit den Fließtexten, Jonathan?

# Jonathan

## Data Extraction

**Folie – Blick in die Maschine (Agenda)**
Gute Frage, Matthias! Auch die unstrukturierten Quellen, die Wikipedia- und Spiegel-Online-Artikel, enthalten wertvolle Informationen über das deutsche Unternehmensnetzwerk.

**Folie – Blick in die Maschine mit Blick in die Data Extraction**
Um auch diese Informationen in den Netzwerkgraphen einfließen zu lassen, nutzen wir als zweite Komponente die Data Extraction.

**Folie – Blick in die Data Extraction mit Beispielsatz**
In die Data-Extraction-Komponente wird nun ein deutschsprachiger Fließtext eingegeben [zeigen]. Bedenken Sie bitte, von welcher Komplexität die deutsche Sprache ist. Für eine Maschine ist sie ohne weiteres nicht verwertbar. Als einfaches Beispiel könnte dieser den folgenden Satz enthalten: „Bosch liefert Servomotoren an die DB."

**Folie – Blick in die Data Extraction mit NER und NEL**
Unsere Software erkennt, dass „Bosch" und „DB" Unternehmensnamen sind. Aus dem Kontext erschließt sie, dass [Animation] „Bosch" die „Robert Bosch GmbH" und „DB" die "Deutsche Bahn AG" aus unserem Netzwerkgraphen bezeichnen. Jetzt analysiert sie die Satzstruktur und ermittelt die Beziehung zwischen beiden Unternehmen. Hierbei kommen aufwändige linguistische Verfahren zur Anwendung. [Animation] Die Robert Bosch GmbH beliefert die Deutsche Bahn AG. Diese Beziehung wird nun in den bestehenden Netzwerkgraphen eingefügt. [Animation, zeigen] Durch die parallele Ausführung auf mehreren Rechnern können wir über 100.000 Texte pro Stunde verarbeiten.

**Folie – Blick in die Maschine (Agenda)**
Unsere Software ergänzt den Netzwerkgraphen um Unternehmensbeziehungen, die in Fließtexten beschrieben werden.

# Data Cockpit

**Folie – Blick in das Data Cockpit**
Zur Qualitätssicherung benötigt ein Risikomanager eine Anzeige zur Überwachung der Datenverarbeitung. Dazu haben wir eine Browseranwendung entwickelt - das Data Cockpit. Es bietet die Möglichkeit, [Einblendung] die Software zu steuern, [Einblendung] statistische Analysen auf den Ergebnissen auszuführen und [Einblendung] Fehler aus den Quellen zu korrigieren.
[Umblendung] Der Netzwerkgraph lässt sich hierzu in übersichtlicher Weise betrachten.

# Schluss

**Folie – Blick in die Maschine (Agenda)**
Unsere Software erkennt unter menschlicher Aufsicht Unternehmen und deren Beziehungen.

**Folie – Resultierender Netzwerkgraph**
Daraus baut sie den Netzwerkgraphen der deutschen Unternehmenslandschaft [Animation]. Ein Risikomanager hat nun die Möglichkeit, das Netzwerk zu untersuchen, um einen tieferen Einblick zu erlangen. Das Ganze funktioniert mit hilfe einer weiteren Browseranwendung:

**Folie – Webapp Demo**
[spreche parallel zum Screencast]
Sucht man nach einem bestimmten Unternehmen, sieht man all seine Beziehungen zu anderen Unternehmen. Das Netzwerk enthält sogar Städte, Länder und Wirtschaftssektoren. Die Art der Beziehungen offenbart sich durch Hovern über die Pfeile. Um komplexere Zusammenhänge zu erkennen, klickt man auf weitere Unternehmen und sieht wiederum deren Beziehungen.
Das dichte Geflecht deutscher Unternehmen ist nun visuell erkennbar und erkundbar.
Nun hat die Commerzbank ein cleveres und schnelles Werkzeug zur Ergänzung des Risikomanagements.

**Folie – Schlussfolie**
Wir möchten uns recht herzlich bei unseren Partnern von der Commerzbank, Oliver Maspfuhl und Dirk Thomas für Ihr großes Interesse an diesem Projekt bedanken. Ein Dankeschön geht natürlich auch an Prof. Naumann, Toni Grütze und Michael Loster, die uns hingebungsvoll betreut haben.