Verein „Sicherung des Friedens“

Jahresthema 2018 „Gestern gehörten meine Daten mir - gehören sie morgen meinen Feinden?“

Erster Vortrag des Jahres 2018 am 20. Februar 2018 in der Hanns-Seidel-Stiftung

PROF. EM. DR. MANFRED BROY, FAKULTÄT FÜR INFORMATIK, TU MÜNCHEN

DIGITALES DATENMANAGEMENT – POTENZIALE UND RISIKEN

Digitalisierung – was ist das? Viele, auch in der Politik, haben keine klare Vorstellung, wird oft mit Breitbandausbau gleichgesetzt (Der Begriff kommt im aktuellen Koalitionsvertrag 400mal vor). Es ist der Übergang von analogen zu digitalen Informationen; vieles wird heutzutage auch digital gespeichert (Fotos, Musik etc.).

Heute wird unter Digitalisierung die schnelle Veränderung in Wirtschaft und Gesellschaft durch die Nutzung digitaler Technologie verstanden.
Früher Informationssammlung auf diese Weise: beginnend mit dem Notieren dessen, was an Lebensmitteln vorhanden war, über die Tontafel, das Buch, den Buchdruck - damit war Information erstmals allgemein zugänglich.
Heute werden Informationen in der Regel digital gespeichert; es wird geschätzt, dass es 2002 erstmals mehr digital als analog weltweit gespeicherte Information gab und 2007 bereits 94 % der weltweiten technologischen Informationskapazität digital war, nach lediglich 3% im Jahr 1993.

Was in jeder Minute passiert auf Google, Facebook und Co. ist eindrucksvoll:

- Mehr als 4 Millionen Anfragen auf Google;
- 100 Stunden neue Videos hochgeladen auf Youtube;
- 2,5 Millionen neue Posts auf Facebook;

- 571 neue Websites generiert;
- 204 Millionen E-Mails gesendet;
- sieben neue Einträge auf Wikipedia erstellt;
- 300.000 Tweets auf Twitter und 220.000 neue Fotos auf Instagram.

Autonome Fahrzeuge sammeln momentan alle Daten, was pro Stunde 4.000 GB an Daten bedeutet – jedes Fahrzeug generiert so viele Daten wie 3.000 Menschen.

Digitalisierung und Vernetzung: Das Wachstum der Datenmengen steigt im Zeitverlauf, die Datenmenge verdoppelt sich alle zwei Jahre. Das Datenvolumen nimmt exponentiell zu mit dem Ausmaß der Vernetzung, dem Internet der Dinge (Internet of Things, IoT): 2012 waren es 2,8 Zettabyte, die Prognosen für 2015 bzw. 2020 sind 12 bzw. 40 Zettabyte, wobei letztere Zahl die 57fache Menge der Sandkörner an allen Stränden der Welt bedeutet – ein Zettabyte ist eine Zahl mit 21 Nullen oder eine Milliarde Terabyte.

Formen von Daten:

- Zahlen(pakete), von Sensoren abgelesen (Geschwindigkeit im PKW);
- Strukturierte Daten in Registern (Adressensätze);
- Sprache: gesprochen – Umwandeln in Schrift;
- Schrift: Interpretation – Umwandeln in Bedeutung;
- Bilder: Analyse – zeigt Röntgenaufnahme einen Krebs? Interpretation – ist Präsident Trump auf dem Bild zu sehen?
- Videos: bewegte Bilder.

Wichtig: wie fallen die Daten an: in Datenbeständen oder in Echtzeit? Wie viel Zeit bleibt für die Datenanalyse? *Besonderes Thema: Mehrere große Datenbestände zu einander in Beziehung setzen.*

Data Analytics – was ist das?

Der Prozess der Gewinnung, Inspektion, Säuberung, Transformation und Modellierung von Daten und Datenbeständen mit dem Ziel, nützliche

Informationen aufzufinden, Schlussfolgerungen zu ziehen und Entscheidungen zu unterstützen. Beispiele: Aktienkurse, Patienten-, Energieverbrauchsdaten, Nebenwirkungen von Drogen …

Data is the new gold! (Open Data Initiative, Europäische Kommission)

Die Rolle der Algorithmen: Es sind nicht die Daten, es sind die Algorithmen, die aus den Daten Erkenntnisse und Entscheidungen ableiten: Data Analytics.

Welche Fragen lassen sich mit diesen Methoden beantworten?

- Fragen allgemeiner Natur, Bsp.: Erzeugt Glyphosat Krebs?
- Spezifische Fragen:Zeigt das Röntgenbild des Patienten X ein Karzinom? Wie hoch ist das Risiko einer Kreditvergabe oder eines Versicherungsbetrugs für Kunde Y? Zeigt das Video einen Elefanten?
- Fragen mit Reaktion: Steht ein Auffahren des LKW auf ein Stauende bevor? Ist ein Tsunami zu erwarten?
- Fragen zur Prädiktion: Steigen die Aktienkurse morgen? Entwicklung des Energiebedarfs in drei Stunden? Wettervorhersage
- Fragen zum Verkehr: MVV-Nutzung - wie viele Fahrgäste fahren womit wohin? Individualverkehr: welches Verkehrsmittel und wohin? Motive für Mobilitätsaktivität? Auswirkungen eines U-Bahn-Neubaus oder der Preisgestaltung im ÖPNV?
- Fragen zum Wohnen: Wo ist der höchste Wohnraumbedarf?
- Querfragen: Welche Baumaßnahmen haben welchen Einfluss auf welche Mobilitätsbedürfnisse?

Welche Aufgaben lassen sich erledigen?

- Umformen von Sprache in Schrift,
- Erkennen des Inhalts von Texten,
- Erkennen von Autokennzeichen oder Gesichtern…,
- Erkennen und Vorhersage von Staus,
- Erkennen von Szenarien (Bsp. Umfeldmodell für autonome Fahrzeuge),
- Analyse des Kaufverhaltens einzelner Personen oder von Personengruppen,

- Erkennen des Zusammenhangs von Krebs mit genetischer Veranlagung,
- Preisgestaltung von Flügen,
- Vorhersage von Wartungserfordernissen (Bsp.: wann muss der Generator ausgetauscht werden?),
- Einschätzen der politischen Gesinnung einer Person anhand ihrer digitalen Spuren,
- Einschätzen, ob eine Frau schwanger ist, ebenfalls anhand ihrer digitalen Spuren,
- Erkennen von Gefühlen und Gemütszuständen von Menschen anhand ihrer Haltung/ihres Gesichtsausdrucks (Lügendetektor),
- Crime Prediction: wie hoch ist die Wahrscheinlichkeit, dass Person X ein Verbrechen begeht?

<u>Medien – digitale Erfassung von Nutzerverhalten</u>

- Welche Berichte interessieren Person X (personalisierte Medieninhalte),
- Wie hoch ist der Wahrheitsgehalt einer Meldung?
- Welche Meldung findet das höchste Interesse (wie viele Personen (clicks) erreicht eine Meldung)?
- Wann ist Person X besonders empfänglich für Werbung?
- Wie ist das Medienangebot am erfolgreichsten zu gestalten?
- Wie viele Personen sind bereits für welchen Inhalte zu bezahlen?
- Wie ist eine Meldung zu gestalten, damit sie den größten Eindruck hinterlässt?

<u>Data Science – Data Mining</u> verwendet

Theorien und Techniken aus Informatik, Mathematik, Statistik, Wahrscheinlichkeitsmodelle, Maschinenlernen, statistischem Lernen, Programmierung und Datentechnik, Mustererkennung, Prognostik, Modellierung (von Unsicherheiten) und Datenhaltung – eine *Trilogie aus Datenerfassung, Datenmodellierung und –analyse und Entscheidungsfindung*

<u>Big Data – was ist das?</u>

Big Data (deutsch: Massendaten) sind große Datenmengen, die sehr schnell in Echtzeit anfallen und zu groß oder zu komplex sind oder sich zu schnell ändern, als dass sie ein Mensch (mit manuellen und klassischen Methoden der Datenverarbeitung) auswerten könnte.

Another Definition of Big Data: „Big Data" refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze (McKinsey Global Institute)

How Big ist Big Data?

1 Petabyte = 1.000 Terabytes (TB) = 10^{15} Bytes

1Zettabyte = 1.000.000.000.000.000.000.000 Bytes = 10^{21} Bytes
Man stelle sich vor, jeder der 320.590.000 US-Amerikaner macht jeden Tag des Monats jede Sekunde ein Foto: alle diese Fotos zusammen ergeben ei Zettabyte! (Big Data: Seizing Opportunities, Preserving Values; Executive Office of the President, May 2014 – The White House, Washington.)

Big Data als Thema: Was ist heute anders?

- Riesige Datenbestände in digitaler Repräsentation (Daten im Internet, Nutzerdaten, Sensordaten),
- Schnellere und leistungsfähigere Rechner,
- Bessere Methoden:
    - in Memory Datenbanken: zu verarbeitende Daten sind im Arbeitsspeicher des Rechners,
    - Hadoop: Datenmengen und ihre Bearbeitung wird auf mehrere Rechner gleichzeitig aufgeteilt,
    - Machine Learning: aus großen Datenbeständen („Trainingsmengen") werden Analysealgorithmen durch Parameterausbildung in neuronalen Netzen geschaffen.
- Mooresches Gesetz (erfand die integrierte Schaltung).

Im Zentrum: Strukturierte Informationen („Wissen") extrahieren aus unstrukturierten, ungenauen, teilweise fehlerhaften und unvollständigen Datenbeständen.
Beispiel: ein Modell aus den Daten der Kfz-Versicherung entwickeln zur Ermitteln des Risikos des Versicherungsnehmers.

Zwei Vorgehensweisen:

- Manuelle Analyse und Erstellen eines Modells, Hypothese mit anschließender Bestätigung/Widerlegung *oder*
- Modell aus den Daten automatisch extrahieren.

Combine linguistics, statistics, and logical reasoning: harder than for „ordinary" relations.

Klassisches Beispiel: Suche im Internet

Datenbestand: Daten im Internet auf Websites

- Finden von Informationen: Anfragebeispiel Präsident der USA?
- Beantworten von Fragen: Wer war Steve Jobs?

Suchmaschinen wie Google stellen Information über die Website zusammen (Webcrawler), suchen auf extrahierten Daten, nicht auf der (gesuchten) Seite; strukturieren die Information für schnelle Suche, geben auf Anfrage Listen von Suchergebnissen aus, die in einer Reihenfolge angeordnet sind (Page Rank).

Information Retrieval

- Informationsexplosion im Internet,
- Ranking von Dokumenten zum Finden relevanter Information,
- Ähnlichkeit von Dokumenten (Bsp. Plagiate in Dissertationen) erkennen.

Beispiel:

- Struktur in Datenbeständen erkennen: ein Fahrzeug hat einen Motor und ein Fahrwerk.
- Zusammenhänge aus Datenbeständen lernen: der Zug Montagmorgen von Passau nach München ist meist verspätet; der Stau beginnt am Freitag um 15 Uhr.
- Regeln aus Datenbeständen lernen: wenn es in der Nacht regnet, gibt es mehr Stau.
- Visualisierung: Stelle dar, wie sich die Stausituation in München in den letzten 20 Jahren entwickelt hat.

Lernende Systeme

- Sogenannte lernende Systeme sind inzwischen (u.a.) leistungsfähige Systeme der Datenanalyse.
- Dramatische Fortschritte in den letzten Jahren durch tiefe neuronale Netze (DNNs), möglich aufgrund höherer Datenmenge, schnellere Hardware, bessere Algorithmen.
- Erfolge:

    - Wahrnehmung: Sehen. Sprache …
    - Spiele: Schach, Go (was beides besser von der Maschine gespielt wird), einfache Videospiele …
    - Einige komplexe Tasks: grammatische Strukturen bei Übersetzung siehe Google-Translator, der nicht mehr analytisch arbeitet, sondern durch Maschinenlernen (ein Lernverfahren, bei dem die Maschine mit Tausenden von Sätzen gefüttert wurde) , autonomes Fahren (ist heute halbwegs auf Autobahnen möglich) …
- Herausforderungen:
    - Verständnis und Erklärbarkeit der gelernten Netze,
    - Methoden zur Validierung und Verifikation,
    - Modularität und Integration mit klassischen Techniken der künstlichen Intelligenz (KI)

Tiefe Datenanalyse

Konzentration auf semantischer Interpretation unstrukturierter Daten: Datamining, Textmining, Informationsextraktion, Automatisierung der Wissensextraktion.

Big Data Challenges – the Big 5s

- Volume (Datenumfang),
- Velocity (die Geschwindigkeit, mit der Daten generiert werden),
- Variety (die unterschiedlichen Arten von Daten),
- Veracity (die vertrauenswürdigkeit von Daten),
- Value (der Wert der enthaltenen Information).

Predictive Analysis: die Zukunft vorhersagen

Modeling, machine learning, statistical analysis, big data are often thrown together in the hopes of predicting future events and behaviors.

Bsp.: Vorhersagen von Energiebedarf (Kosten und Verbrauch) in den nächsten fünf Stunden.

Forbes-Umfrage (Forrester): 10 Vorhersagen für Künstliche Intelligenz, Big Data und Analyse für 2018: 70 % der Unternehmen erwarten die Implementierung von KI innerhalb eines Jahres, gegenüber 40 % 2016 und 51 % 2017. Data Engineer wird der neue „heiße“ Jobtitel.

Ist KI künstliche „Intelligenz“? Ist maschinelles Lernen „Lernen“?

Die Begriffe täuschen!

- KI-Systeme (Bsp. Sprache in Schrift. Schrift in Semantik) arbeiten völlig anders als der Mensch – sie sind nicht „intelligent“ wie Menschen, sondern lösen bestimmte anspruchsvolle Aufgaben – jedoch mit ganz anderen Mitteln;
- Lernende Systeme („Machine Learning“) lernen nicht wie Menschen (erkennen aus wenigen Beispielen Regeln und Assoziationen), sondern arbeiten mit „Brute Force“: aus Millionen von Bildern als Trainingsset werden in einem „neuronalen Netz“ (der Begriff bedeutet etwas anderes als die Neuronen im Hirn!) die Parameter so gesetzt, dass erkannt wird, auf welchem Bild Katzen sind.

<u>BR: Künstliche Intelligenz: Computer lesen jetzt besser als Menschen</u>

- Mitte 2016: der KI-Bewerber einer Universität aus Singapur schaffte 51 % der Fragen richtig zu beantworten – der menschliche Bewerber schaffte 82,3 %.
- KI macht momentan große Sprünge: Programm des chinesischen Konzerns Alibaba und von Microsoft haben Punktestand der menschlichen Testperson geschlagen.
- Maschinen können jetzt Fragen wie „Was ist die Ursache für Regen?“ mit hoher Genauigkeit beantworten.

<u>The Instability of neural networks</u>: Eine kleine Störung beim Input kann große Auswirkungen haben – beispielsweise ein Dreckfleck oberhalb der Ziffer „60“ einer Geschwindigkeitsbegrenzung wird fälschlich als „80“ erkannt.

<u>Der Verbraucher in den digitalen Medien</u>

- Hohe Transparenz: durch Bewertungsportale und Vergleich von Angeboten;
- Sammeln von Daten über den Verbraucher ist in den digitalen Medien nahezu unbegrenzt möglich:
    - Profiling: Interesse an welchen Waren/Dienstleistungen?
    - Situation und Verhalten! Stimmung?
    - Verhältnisse des Verbrauchers?
- D.h. der Verbraucher wird durchschaubar und manipulierbar:
    - Preise unterschiedlich für Flüge etc.,
    - gezielte Angebote – „zur richtigen Zeit, im richtigen Moment“,
    - Diskriminierung!

<u>Der überwachte Bürger – ein Beispiel</u>

- *Harmlos?*
  Es ist technisch möglich, Autofahrer digital zu erfassen (Geschwindigkeit, Route, etc., auch Alkoholisierung oder Müdigkeit) –

was es erlaubt, Maut und Verstöße gegen Verkehrsregeln umfassend zu registrieren.

- *Die ethische Frage*:
  Haben Menschen ein Recht auf Unbeobachtetsein und die Freiheit, gewisse Risiken einzugehen (inklusive Verstoß gegen Regeln oder Gesetze mit der Chance, nicht belangt zu werden?)
  Wo sind die Grenzen?
- *Diverse Möglichkeiten, Daten über Bürger zu sammeln,*
  - sind nicht nur für Unternehmen von großem Interesse,
  - sondern werden von totalitären Staaten zunehmend zur Überwachung eingesetzt.
- Totalitäre Staaten haben mit den verfügbaren Daten, die schon heute umfassend über jeden einzelnen anfallen, eine Fülle von Möglichkeiten zur Überwachung und Steuerung des Verhaltens ihrer Bürger.
- China: mehrere konkurrierende Scoring-Systeme (digitale Reputationssysteme) weisen Bürgern einen „sozialen" Punktestand zu, gespeist aus Online- und Offline-Daten.
  - Gespeist aus Zahlungsmoral, politischer Aktivität, dem Punktestand seiner Bekannten ergibt sich ein Wert.
  - Der Wert steigt oder sinkt abhängig vom Verhalten und bestimmt den Zugang zu Bildung, Kredit und Konsum.
  - Die Systeme bleiben für den betroffenen weitgehend intransparent.
  - Ein spielerischer Ansatz ist besser zur Durchsetzung konformen Verhaltens als Drohen und Strafen.
  - Eingeflossen sind Erfahrungen aus PC-Spielen.
- Digitale Reputationssysteme schränken Fähigkeit und Willen der Betroffenen zu Protest gegen Ungerechtigkeit ein (Social Cooling).
- „Social Cooling" – eine direkte Auswirkung der realen oder vermeintlichen Überwachung. Wer sich überwacht fühlt, unterlässt Handlungen, die ein Überwacher missverstehen oder missbilligen könnte.
- Unbewusste und bewusste Verhaltensveränderung durch „Social Cooling" wird in China bereits praktiziert:
  - Bürger erhalten eine von der Regierung vorgeschriebene „soziale Kreditwürdigkeit" – abhängig von ihrem Verhalten, ihren Kriminalakten, ihren Aussagen in sozialen Medien, ihren Einkäufen und auch den Noten ihrer Freunde;
  - was einen subtilen Überwachungsstaat charakterisiert, in dem alle Handlungen einer im- oder expliziten Kontrolle unterliegen.

<u>Der gläserne Bürger – politische Überwachung</u>:

- Auch die politische Einstellung lässt sich erfassen.
- Der totale Staat kann über die digitalen Möglichkeiten seine Bürger in einem Umfang überwachen – und manipulieren, welches alles bisher Dagewesene dramatisch übertrifft.

<u>Die Herausforderungen: Wirtschaft – Recht – Ethik</u>

- Datenanalysetechniken schnell für die Wirtschaft verfügbar machen: Kompetenzzentren!
- Klären: Wem gehören welche Daten? (Wie sind die Daten zu teilen?)
- Klären: Wo sind die Grenzen der Datenanalysetechniken – Privatheit! Freiheitlich-demokratische Grundordnung!

<u>Die ethische Dimension: Werte</u>

Risiken:

- Werden Menschen auf ihre Daten reduziert?
- Werden Menschen: vermessbar, manipulierbar, den Entscheidungen von Maschinen unterworfen, ihrer Privatheit beraubt?
- Wird ihre Identität gestohlen?
- Werden Daten manipuliert und gefälscht?

Potenziale:

Können Daten mehr Sicherheit, Effizienz, Komfort, Erkenntnis
in Gebieten wie
Medizin und Gesundheit, Verkehr, Finanzwesen, Politik etc.
schaffen?

EVA DITTRICH, ASS. JUR.