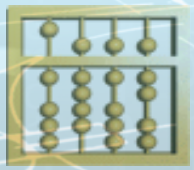
Big Data - Fluch oder Segen

Manfred Broy

Technische Universität
München
Institut für Informatik



ZENTRUM
DIGITALISIERUNG.
BAYERN

Impressive numbers: New data per minute

more than 4 Mio queries	Google	100 hours new video material uploaded	YouTube
2.5 Mio new posts	facebook	571 new websites generated	http://
more than 204 Mio emails sent	@	300,000 tweets generated	twitter
220,000 new pictures uploaded	Instagram	7 new Wikipedia articles	

Daten aus Cyber-physikalischen Systemen

- Der Output eines autonomen Fahrzeugs: 4.000 GB Daten. Pro Stunde
- Intel (Lt. CEO Brian) geht man von etwa 40 Terabyte Daten aus, die pro 8 Stunden autonomer Fahrzeit anfallen.
  - bei einer Stunde durchschnittlicher Fahrweise etwa 4.000 Gigabyte.
- Bei den Kameras rechnet man mit 20 bis 40 Mbps beim Radar mit 10 bis 100 Kbps.
- Jedes autonome Fahrzeug generiert in etwa soviel Daten wie 3.000 Menschen.
  - Eine Million Autos emitieren Daten, die der typischen Datennutzung einer Gruppe von 3 Milliarden Menschen entspricht.

Digitalisierung und Vernetzung

Das Wachstum der Datenmengen im Zeitverlauf

Datenvolumen [in byte]

Zetta

Exa

Peta
Tera
Giga
Mega

Datenvolumen 2012: 2,8 Zettabyte
2.800.000.000.000.000.000.000 Byte

Prognosen für
Datenvolumen 2015: 12 Zettabyte
Datenvolumen 2020: 40 Zettabyte

1 Zettabyte ≙ 1 Mrd. Terabyte

➡ **die erzeugten Datenmengen steigen exponentiell mit dem Ausmaß der Vernetzung.**

➡ **Datenmengen verdoppeln sich alle 2 Jahre**

Zum Vergleich:
40 Zettabytes ≙

Menge an Sandkörnern aller Strände dieser Welt

IoT
Industrie 4.0
Cloud Computing
Big Data / Datability
Mobile

2014
2007
2006
2004
1998
1994
1991
1981

www

Buchdruck

Erster Computer
Transistor

Internet

1438 1878 1896 1926 1938 1947 1969

Quellen: BITKOM, EMC, 2012.

Wo kommen die Daten her?

- Nutzerinteraktionen
  - Jede Aktion eines Nutzers mit einem digitalen System hinterlässt Spuren
    - Besuch bestimmter Web-Seiten
    - Kauf bestimmter Waren
    - Chats
    - e-Mails
    - Bilder
    - ...
  - Es entsteht eine digitale Spur
- Sensoren in Cyber-Physical Systems
  - Die Sensoren eines Smartphones können vielfältige personenbezogene Daten erfassen (Bewegung, Position)
  - In den unterschiedlichsten systemen werden Daten erfasst

Formen von Daten

- Zahlen und Zahlenpakete
  - von Sensoren (Beispiel Geschwindigkeit)
- Strukturierte Daten
  - Adressensätze
- Sprache
  - gesprochen – Umwandlung in Schrift
- Schrift
  - Interpretation – Umwandlung in Bedeutung
- Bilder
  - Analyse – zeigt die Röntgenaufnahme eine Form von Krebs
  - Interpretation – ist Trump auf dem Bild zu sehen
- Videos – bewegte Bilder

Wichtig: Wie fallen die Daten an:
- in Datenbeständen
- in Echtzeit

und wieviel 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,
- Modellierung

von Daten und Datenbeständen mit dem Ziel

- nützliche Information aufzufinden
- Schlussfolgerungen zu ziehen
- Entscheidungen zu unterstützen

Beispiele: Aktienkurse, Patientendaten, Energieverbrauchsdaten, Nebenwirkungen von Drogen ...

> *"Data is the new gold"*
>
> (Open Data Initiative, European Commission. aim at opening up Public Sector Information).

Die Rolle der Algorithmen

- Es sind nicht nur die Daten, es sind die Algorithmen, die aus den Daten Erkenntnisse und Entscheidungen ableiten:

Data Analytics

- Aber es sind die neuen Möglichkeiten, an Massen von Daten zu kommen
  - Die Clicks bei der Bedienung eines Key Boards
  - Die Eingaben ins Internet
  - Die Sensoren der eingebetteten Systeme
  - ...

Welche Fragen lassen sich mit diesen Methoden beantworten?

- Fragen allgemeiner Natur:
  - Beispiel: Erzeugt Glyphosat Krebs?
- Spezifische Fragen:
  - Zeigt das Röntgenbild von Patient X ein Karzinom?
  - Wie hoch ist das Risiko einer Kreditvergabe/eines Versicherungsvertrags für Kunde Y?
  - Zeigt das Video einen Elefanten?
- Fragen mit Reaktion:
  - Steht ein Aufprall des LKW auf ein Stauende bevor?
  - Ist ein Tsunami zu erwarten?
- Fragen zur Prädiktion:
  - Steigen die Aktienkurse bis morgen?
  - Wie entwickelt sich der Energiebedarf in den nächsten 3 Stunden?
  - Wie wird das Wetter?

Welche Fragen lassen sich mit diesen Methoden beantworten?

- Fragen zum Verkehr:
  - Wie nutzen die Fahrgäste den MVV?
  - Von welchen Orten zu welchen Orten fahren Menschen mit welchen Verkehrsmitteln im Individualverkehr?
  - Was sind die Motive für Mobilitätsaktivität?
  - Wie wirkt sich es aus, wenn eine neue U-Bahn gebaut wird?
  - Wie wirken sich die Preisgestaltung auf die Nutzung des ÖNVP aus?
- Fragen zum Wohnen:
  - Wo ist der höchste Wohnraumbedarf?
- Querfragen:
  - Welche Baumaßnahmen haben welchen Einfluss auf welche Mobilitätsbedürfnisse?

Welche Aufgaben lassen sich erledigen?

- Umformung von Sprache in Schrift
- Erkennen des Inhalts von Texten
- Erkennen von
  - Autokennzeichen
  - Gesichtern
  - ...
- Erkennen und Vorhersage von Stau
- Erkennen eines Szenarios
  - Beispiel: Umfeldmodell für autonome Fahrzeuge
- Analyse des Kaufverhaltens
  - Einzelner Personen
  - Von Personengruppen
- Erkennen Zusammenhang Krebs mit genetischer Veranlagung

Welche Aufgaben lassen sich erledigen?

- Preisgestaltung von Flügen
- Vorhersage von Wartungserfordernissen:
  - Wann muss der Generator ausgetauscht werden?
- Einschätzung der politischen Gesinnung anhand der digitalen Spuren einer Person.
- Einschätzung, ob eine Frau schwanger ist, anhand der digitalen Spuren.
- Erkennen der Gefühle und Gemütszustände von Menschen anhand des Haltung oder des Gesichtsausdrucks (Beispiel Lügendetektor).
  - Vorhersage von Handlungen: Wie hoch ist der Preis, den Person Z bereit ist zu zahlen.
- Crime Prediction: Wie hoch ist die Wahrscheinlichkeit, das Person X ein Verbrechen begeht?

Medien – digitale Erfassung von Nutzerverhalten

- Welche Berichte interessieren Person X?
  - Personalisierte Medieninhalte.
- Wie hoch ist der Wahrheitsgehalt einer Meldung?
- Welche Meldung findet das höchste Interesse?
  - Wieviele Personen (clicks) erreicht eine Meldung?
- Wann ist Person X für Werbung besonders empfänglich?
- Wie ist das Medienangebot am erfolgreichsten zu gestalten?
- Wieviel sind Personen bereit für welche Medieninhalte zu bezahlen?
- Wie ist eine Meldung zu gestalten, dass sie den höchsten Eindruck hinterlässt?

Data Science – Data Mining

... verwendet

- Techniken und Theorien aus den Fächern Mathematik, Statistik und Informatik
- Wahrscheinlichkeitsmodelle
- Maschinenlernen,
- statistisches Lernens
- Programmierung und Datentechnik
- Mustererkennung
- Prognostik
- Modellierung
  - von Unsicherheiten
- Datenhaltung

Trilogie aus
- Datenerfassung,
- Datenmodellierung und -analyse
- Entscheidungsfindung

Big Data – was ist das?

Big Data (Deutsch: Massendaten): Datenmengen, die

- sehr schnell in Echtzeit anfallen,
- zu groß oder zu komplex sind oder
- sich zu schnell ändern,

um sie mit manuellen und klassischen Methoden der Datenverarbeitung auszuwerten.

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 is Big Data?

1 ***petabyte*** is 1,000 terabytes (TB)= 10^{15}bytes

1 ***zettabyte*** is 1,000 000,000,000,000,000,000 bytes == 10^{21}bytes

"***Imagine that every person in the United States (***320,590,000) ***took a digital photo every second of every day for over a month. All of those photos put together would equal about one zettabyte***" (*)

(*) *BIG DATA: SEIZING OPPORTUNITIES, PRESERVING VALUES*
Executive Office of the President, MAY 2014 -The White House, Washington.

Big Data -

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 befinden sich im Arbeitsspeicher des Rechners
  - Hadoop
    Datenmengen und ihre Bearbeitung werden auf mehrere Rechner gleichzeitig aufgeteilt
  - Machine learning
    Aus großen Datenbeständen („Traingsmengen“) werden Analyse–algorithmen durch Parameterausbildung in Neuronalen Netzen geschaffen

Im Zentrum

- Aus unstrukturierten, ungenauen, teilweise fehlerhaften und unvollständigen Datenbeständen
- strukturierte Informationen („Wissen“) zu extrahieren.

Beispiel: Aus den Daten der Kfz-Versicherung ein Modell entwickeln, welcher Versicherungsnehmer welches Risiko bedeutet.

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

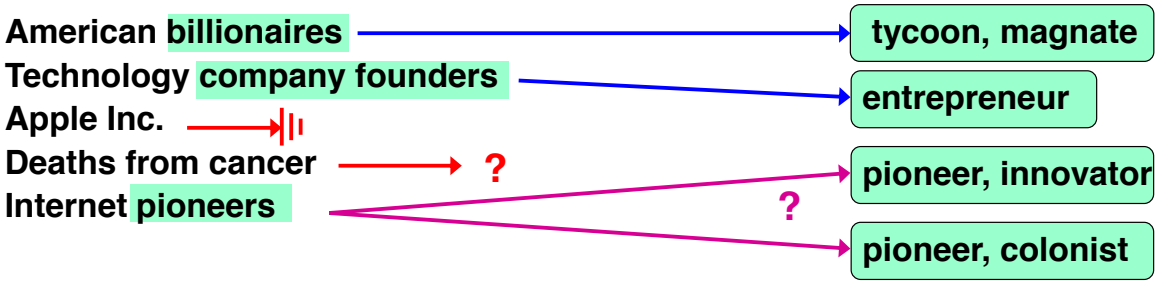
Zwei Vorgehensweisen:

- Manuelle Analyse und Erstellung eines Modells, Hypothese mit anschließender Bestätigung oder Widerlegung
- Modell aus den Daten automatisch extrahieren

Klassisches Beispiel: Suche im Internet

Datenbestand: Daten im Internet auf web-Seiten

- Finden von Informationen
  - ◇ Anfragebeispiel: Präsident der Vereinigten Staaten
- Beantwortung von Fragen
  - ◇ Wer war Steve Jobs?



Suchmaschinen:

- stellen Information über die web-Seiten zusammen (Web-Crawler)
- strukturieren die Information für die schnelle Suche
- geben auf Anfragen Listen von Suchergebnissen aus, die in einer Reihenfolge angeordnet sind (Page Rank)

Beispiel

- Struktur in Datenbeständen erkennen
  - Ein Fahrzeug hat einen Motor und ein Fahrwerk
- Zusammenhänge aus Datenbeständen lernen
  - Der Zug Montag morgen von Passau nach München ist meist verspätet
  - Der Stau beginnt am Freitag um 15:00
- 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 Instrumente der Datenanalyse
- Dramatische Fortschritte in den letzten paar Jahren
  - Hauptsächlich durch Tiefe-Neuronale-Netze (DNNs)
  - Möglich, da mehr Daten, schnellere HW, bessere Algorithmen
- Erfolge
  - Wahrnehmung: Sehen, Sprache, ...
  - Spiele: Schach, Go, einfache Videospiele, ...
  - Einige komplexe Tasks: Übersetzung, autonomes Fahren, ...
- Herausforderungen
  - Verständnis und Erklärbarkeit der gelernten Netze
  - Methoden zur Validierung und Verifikation
  - Modularität und Integration mit klassischen KI-Techniken

Tiefe Datenanalyse

Konzentration auf semantischer Interpretation unstrukturierter Daten

- Datamining
- Textmining
- Informationsextraktion
- Automatisierung der Wissensextraktion

Predictive Analytics: Die Zukunft vorhersagen

- Modeling,
- machine learning,
- statistical analysis
- Big Data

are often thrown together in the hopes of predicting future events and behaviors.

Beispiel:
Sage den Energieverbrauch und die Energiekosten in den nächsten 5 Stunden voraus

Forbes: 10 Predictions For AI, Big Data, And Analytics in 2018

70% of enterprises expect to implement AI over the next 12 months, up from 40% in 2016 and 51% in 2017.

Summary of what Forrester predicts will happen in 2018:

- 25% of enterprises will supplement point-and-click analytics with conversational interfaces.
  - Querying data using natural language and delivering resulting visualizations in real time will become standard features of analytical applications.
- 20% of enterprise will deploy AI to make decisions and provide real-time instructions.
  - AI will suggest what to offer customers, recommend terms to give suppliers, and instruct employees on what to say and do — in real time.

Forbes: 10 Predictions For AI, Big Data, And Analytics in 2018

- AI will erase the boundaries between structured and unstructured data-based insights.
  - The number of global survey respondents at enterprises with more than 100 terabytes of unstructured data has doubled since 2016. Deep learning has made analyzing this type of data more accurate and scalable.
- 33% of enterprises will take their data lakes off life support.
  - Without a clear connection to change-the-business outcomes, many early adopters will pull the funding plug on their data lakes to see if they pay for themselves or die.
- 50% of enterprises will adopt a cloud-first strategy for big data analytics.
  - Forrester expects 50% of enterprises to embrace a public-cloud-first policy in 2018 for data, big data, and analytics, for more control over costs and more flexibility than on-premises software can deliver.

Forbes: 10 Predictions For AI, Big Data, And Analytics in 2018

- 66% of enterprises will deploy insight centers of excellence as a remedy for organizational misalignments.
  - With firms bringing the voice of the customer into every business decision in a unified way, 56% of enterprises already report creating customer insight centers of excellence rather than centralized or purely distributed models to accomplish this.
- The majority of Chief Data Officers (CDOs) will move from defense to offense.
  - Business-oriented CDOs will explore opportunities to innovate with data, either through analytics embedded in internal business processes or through new external data-enabled products and services. In 2018, more than 50% of CDOs will report to the CEO , up from 34% in 2016 and 40% in 2017.

Forbes: 10 Predictions For AI, Big Data, And Analytics in 2018

- Data engineer will become the hot new job title.
  - 13% of data-related job postings on Indeed.com are for data engineers, reflecting the trend of big data initiatives becoming mission-critical and need to provide broader support to the business analyst.
- The insights-as-a-service market will double as insight subscriptions gain traction.
  - 66% of enterprises already outsource between 11% and 75% of their Business Intelligence applications. Forrester predicts that up to 80% of firms will rely on insights service providers for some portion of their insights capabilities in 2018.
- Academia will become the new insights partner for enterprises.
  - And not just academia—new research labs like the nonprofit Open AI help solve the most challenging analytic and AI problems for firms that submit requests.

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

Achtung, die Worte täuschen:

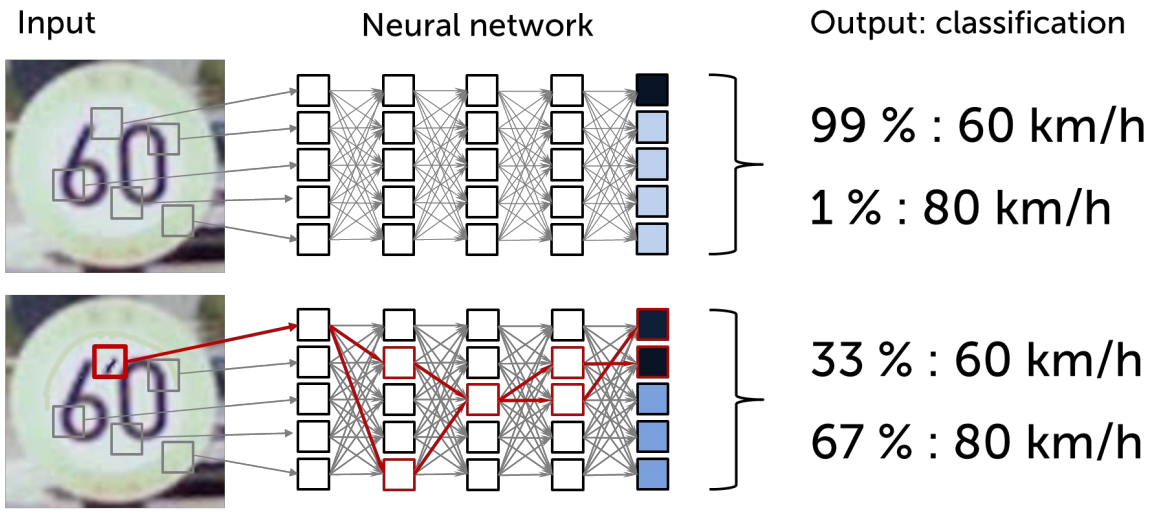
- KI Systeme (Beispiel: Sprache in Schrift, Schrift in Semantik) arbeiten völlig anders als der Mensch – sie sind nicht „intelligent" wie Menschen sondern lösen bestimmte anspruchsvolle Aufgaben – jedoch mit ganz anderen Mitteln
- Lernende Systeme („Machine Learning") lernen nicht wie Menschen (erkennen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

BR: Künstliche Intelligenz: Computer lesen jetzt besser als Menschen

- Test der Universität Stanford:
  - Zum Leseverständnis sind Fragen zu Wikipedia-Artikeln beantworten, nachdem diese (ein-)"gelesen" wurden.
  - Über 100.000 Verständnisfragen zu rund 500 Wikipedia-Artikeln wurden über Mechanical Turk, die Online-Arbeitsplattform von Amazon, gewonnen.
- **Maschinen, die aufmerksamer lesen als Menschen**
  - Mitte 2016: Der KI-Bewerber von einer Universität aus Singapur schaffte 51 Prozent der Fragen richtig zu beantworten
  - deutlich schlechter als die menschliche Testperson mit 82,3 Prozent
- KI macht gerade Quantensprünge.
  - Heute: Programm des chin. Konzerns Alibaba und eines von Microsoft haben Punktestand der menschlichen Testperson geschlagen.
  - Maschinen können jetzt Fragen wie "Was ist die Ursache für Regen?" mit hoher Genauigkeit beantworten.

Dependable neural networks

Dependable neural networks are crucial for safe and secure autonomous and decision systems



A tiny disturbance in the input causes a significant change in the classification.

Formal verification for artificial neural networks © fortiss GmbH

Der Verbraucher in den digitalen Medien

- Hohe Transparenz für den Verbraucher durch Digitalisierung
  - Bewertungsportale
  - Vergleich von Angeboten
- Die Möglichkeiten des Sammelns von Daten über den Verbraucher in den digitalen Medien nahezu unbegrenzt
  - Interesse an welchen Waren oder Dienstleistungen – Profiling
  - In welcher Situation welches Verhalten
  - Welche Stimmung
  - Aus welchen Verhältnissen stammt der Verbraucher
- Der Verbraucher wird durchschaubar und manipulierbar
  - Preise unterschiedlich (Flüge etc.)
  - Gezielte Angebote („zur richtigen Zeit, im richtigen Moment“)
  - Diskriminierung

Der überwachte Bürger – ein Beispiel

- Harmlos?
  - Es ist technisch möglich, Autofahrer digital zu erfassen (Geschwindigkeit, Route, ..., auch Alkoholisierung oder Müdigkeit)
  - Dies erlaubt es Maut oder Verstöße gegen Verkehrsregeln umfassend zu registrieren
- Die ethische Frage:
  - Haben Menschen ein Recht darauf, unbeobachtet zu sein und die Freiheit gewisse Risiken einzugehen (auch Verstoß gegen Regeln oder gar Gesetze mit der Chance nicht belangt zu werden)?
  - Wo sind die Grenzen?

Der überwachte Bürger

- Diverse Möglichkeiten, Daten über den Bürger zu sammeln,
  - sind nicht nur für Unternehmen von großem Interesse,
  - werden von totalitären Staaten zunehmend zur staatlichen Überwachung eingesetzt.
- Totalitäre Staaten haben mit den schon heute umfassend verfügbaren Daten, die über jeden einzelnen digital anfallen, eine Fülle von Möglichkeiten zur Überwachung und Steuerung.

Der überwachte Bürger

- In China weisen mehrere konkurrierende Scoring-Systeme Bürgern einen „sozialen" Punktestand zu, der sich aus Online- und Offline-Daten über sie speist.
  - Aus Zahlungsmoral, politischer Aktivität, weiteren digitalen Daten und den Punktesständen des Bürgers und auch 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 Einzelnen weitgehend intransparent.
  - Um konformes Verhalten durchzusetzen, ist dabei ein spielerischer Ansatz besser als Drohen und Strafen.
  - Erfahrungen aus Computerspielen flossen in die chinesischen Systeme ein.
- Digitale Reputationssysteme schränken Fähigkeit und Willen der Menschen ein, gegen Ungerechtigkeit zu protestieren.

Der überwachte Bürger

- "Social Cooling" – eine direkte Auswirkung der realen oder vermeintlichen Überwachung.
  - Vorhaben der Bürger, die sich überwacht fühlen, werden nicht mehr in Handlungen umgesetzt - es könnte nicht gut aussehen, für jemand, der einen überwacht.
- Unbewusste und bewusste Verhaltensveränderung durch „Social Cooling" wird in China bereits praktiziert.
- In China erhalten Bürger eine von der Regierung vorgeschriebene "soziale Kreditwürdigkeit".
  - Die zeigt, wie gut sie sich verhalten, und basiert auf Kriminalakten, was sie in den sozialen Medien sagen, was sie kaufen und sogar den Noten ihrer Freunde.
  - Das charakterisiert einen subtilen Überwachungsstaat, in dem alle Handlungen einer impliziten oder expliziten Kontrolle unterliegen.

Der gläserne Bürger - politische Überwachung

- Auch die politische Einstellung lässt sich erfassen
- Der totalitäre Staat kann über die digitalen Möglichkeiten Bürger in einem Umfang überwachen (und manipulieren), der alles dramatisch übertrifft, was bisher möglich war.

Die Herausforderungen: Wirtschaft, Recht, Ethik

- Datenanalysetechniken schnell für die Wirtschaft verfügbar machen!
  - Kompetenzzentren
- Klären: Wem gehören welche Daten?
  - Wie Daten teilen?
- Klären: Wo sind die Grenzen von Datensammelwut und Datenanalysetechniken
  - Privatheit!
  - Freiheitliche Grundordnung

Die ethische Dimension: Werte

Die 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?

Die Potenziale:

Können Daten mehr

- Sicherheit
- Effizienz
- Komfort
- Erkenntnis

in Gebieten wie

- Medizin und Gesundheit
- Verkehr
- Finanzwesen
- Politik
- ...

schaffen