Wie und wo Computer rasant ihr Wissen vermehren

Klaus Birkenbihl, ict-Media GmbH
Wie schnell kommt Cognitive Computing, *11. November 2015, 13.00-19.00 Uhr*
Deutsche medienakademie GmbH in Kooperation mit der Bitkom Akademie
http://www.ict-media.de/cognitive-computing/text.pdf

In den 50er und 60er Jahren des letzten Jahrhunderts hießen Computer noch hochstaplerisch Elektronengehirn. Der Name manifestiert einen Anspruch und ein Ziel: die Funktionen des menschlichen Gehirns sollen durch Computer ergänzt, nachgebildet oder gar übertroffen werden. Wie nahe wir dran sind, und einige der Konsequenzen, davon handelt mein Statement. In meinem Berufsleben hatte ich hin und wieder mit diesem Thema zu tun – das erste mal 1970 während meines Studiums. Die Beschäftigung mit dem Thema war immer geprägt von Faszination, Erwartungen, aber auch Enttäuschungen. Doch jetzt scheint ein großes Zwischenziel erreicht.

Sir Timothy's Semantic Web Szenario

Abbildung 1: Scientific American May 2001

In 2001 erscheint im Scientific American ein Artikel mit etwa folgendem Szenario: „Auf der heimischen Musikanlage rocken die Beatles 'we can work it out'. Ein Anruf kommt für Pete. Petes Schwester ruft von der Arztpraxis aus an. 'Mutter braucht eine Behandlung beim Physiotherapeuten, zweiwöchentlich. Ich lasse meinen Agenten die Termine ausmachen.' Pete erklärt sich sofort bereit Chauffeurdienste zu übernehmen." Was folgt ist die Schilderung, was die Agenten unternehmen, um das Ganze so zu organisieren, dass Kriterien wie 'Lucy oder Pete haben Zeit zu fahren', 'der Therapeut hat eine gute Bewertung', 'Vertrauen in die Bewertungen', 'freie Termine beim Therapeuten', 'Fahrwege und Verkehr auf dem Weg zum Therapeuten' optimiert werden. Die Agenten - versteht sich - sind digitaler Natur und unterhalten sich mit ihren Chefs über 'mobile Web Browser'.

Einer der Autoren dieses Artikels ist Sir Timothy Berners-Lee, der Erfinder des *World Wide Web*, in dessen *World Wide Web Consortium (W3C)* ich mehrere Jahre im Team und im Advisory Board arbeiten durfte. Die technische Infrastruktur, auf deren Basis die Agenten arbeiten, nennt Berners-Lee *Semantic Web*. Einige Details der Geschichte sind heute Alltag - z.B. mobile Browser. Anderes fühlt sich so an, als würde es gerade umgesetzt durch digitale Assistenten. *Siri, Cortana* oder *Google Now* stehen auf dem Sprung unser Leben zu organisieren. Ein Grund, einmal genauer hinzusehen und herauszufinden, wo wir stehen und was da auf uns zu rollt und uns über Chancen und Risiken nachzudenken.

Das Semantic Web

Berners-Lees Szenario ist deshalb so präzise, weil es – wie u.a. das *WWW* – auf vorhandenen Tech-

nologien basiert, diese standardisiert und miteinander zu vernetzt. Die Grundlagen für die Technologien, die beim Semantic Web zum Einsatz kommen, wurden – wie vieles im *Cognitive Computing* – überwiegend in den 80er Jahren des vorigen Jahrhunderts entwickelt. Die Probleme, die sie lösen:

- Bedeutung von Begriffen ausdrücken: die Bedeutung wird durch die Beziehungen (relations), in denen ein Begriff zu anderen Begriffen steht, definiert,
- Die globale Eindeutigkeit von Begriffen herstellen, Synonyme identifizieren,
- Gruppierung und Klassifizierungen von Begriffen durch Ontologien und Taxonomien. Verfahren zur Vernetzung von Taxonomien und Ontologien,
- Wissen repräsentieren: die formale Beschreibung von Zusammenhängen und Schlussfolgerungen.

Wesentlicher Bestandteil des *Semantic Web* sind *Uniform resource Identifiers* (*URIs)*, die die Eindeutigkeit von Begriffen und Relationen herstellen. So wie im *WWW* die *Uniform resource Locators (URLs)* eindeutig Inhalte im WWW und Internet identifizieren, so beschreiben *URI*s zusätzlich eindeutig die Dinge, Ideen und Individuen der realen Welt.

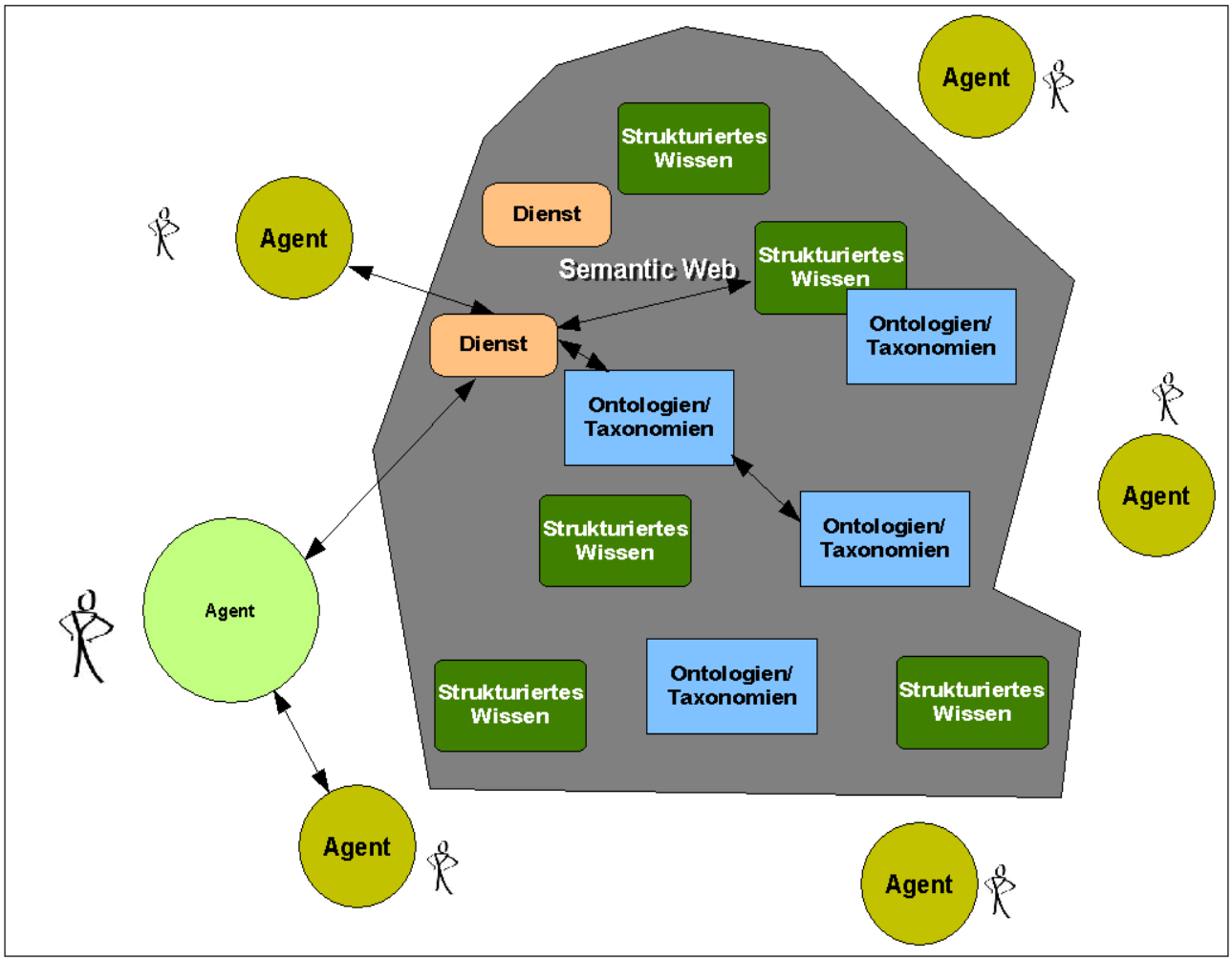

Abbildung 2: Semantic Web

Agenten operieren zur Lösung ihrer Aufgaben auf dieser Infrastruktur, sie kommunizieren miteinan-

der und tauschen die zur Lösung ihrer Aufgaben erforderlichen Daten aus, bilden – z. B. durch Kryptographie gestützte – Vertrauensnetzwerke. Eine vereinfachte technische Struktur des Semantic Web zeigt *Abbildung 2*.

Strukturiertes Wissen erklärt

Strukturiertes Wissen sind Fakten, dargestellt in einer standardisierten Form. Diese Form erlaubt es, sie in Bezug zu anderen Fakten zu setzen. Taxonomien und Ontologien beschreiben so Klassen und Begriffe, über die sich das Wissen definiert. Ähnlich funktioniert unser Wissen. „Caro“ - sagt uns nichts. „Caro ist ein Hund“ ordnet „Caro“ einer Klasse zu, die in unserem Hirn repräsentiert ist. Die Aussage „Caro gehört einer Frau, die Frau wohnt in Köln“ setzt „Caro“ in Relation zu einer Instanz aus „Frau“, die eine „Unterklasse“ von „Mensch“ ist, die eine „Unterklasse“ von „Säuger“ ist, zu der auch „Hund“ gehört. Aber die „Frau“ „wohnt in“ „Köln“, „Köln“ ist eine „Stadt“, in einer „Stadt“ „wohnen“ „Menschen“ … Wir sehen, wie sich in unserem Hirn neue Aussagen schnell mit vorhandenem Wissen verbinden und Sinn machen. Solche Strukturen erlauben es auch, logische Schlüsse zu ziehen. Sollten wir nicht wissen, dass „Caro“ ein „Hund“ ist, so erlaubt uns die Beobachtung „Caro bellt“, das Wissen, dass „wenn *x* bellt ist *x* mit hoher Wahrscheinlichkeit ein Hund“ und „Caro ist ein beliebter Hundename“ zu schließen, dass „Caro“ mit hoher Wahrscheinlichkeit ein Hund ist.

Die Praxis: Linked Open Data

Berners-Lee erwartete, dass sein Szenario in wenigen Jahren umgesetzt würde. Und in der Tat sind einige – teils recht große – auf diesen Technologien basierende Infrastrukturen entstanden. Eine der bekanntesten ist *Linked Open Data (Abbildung 3)*.

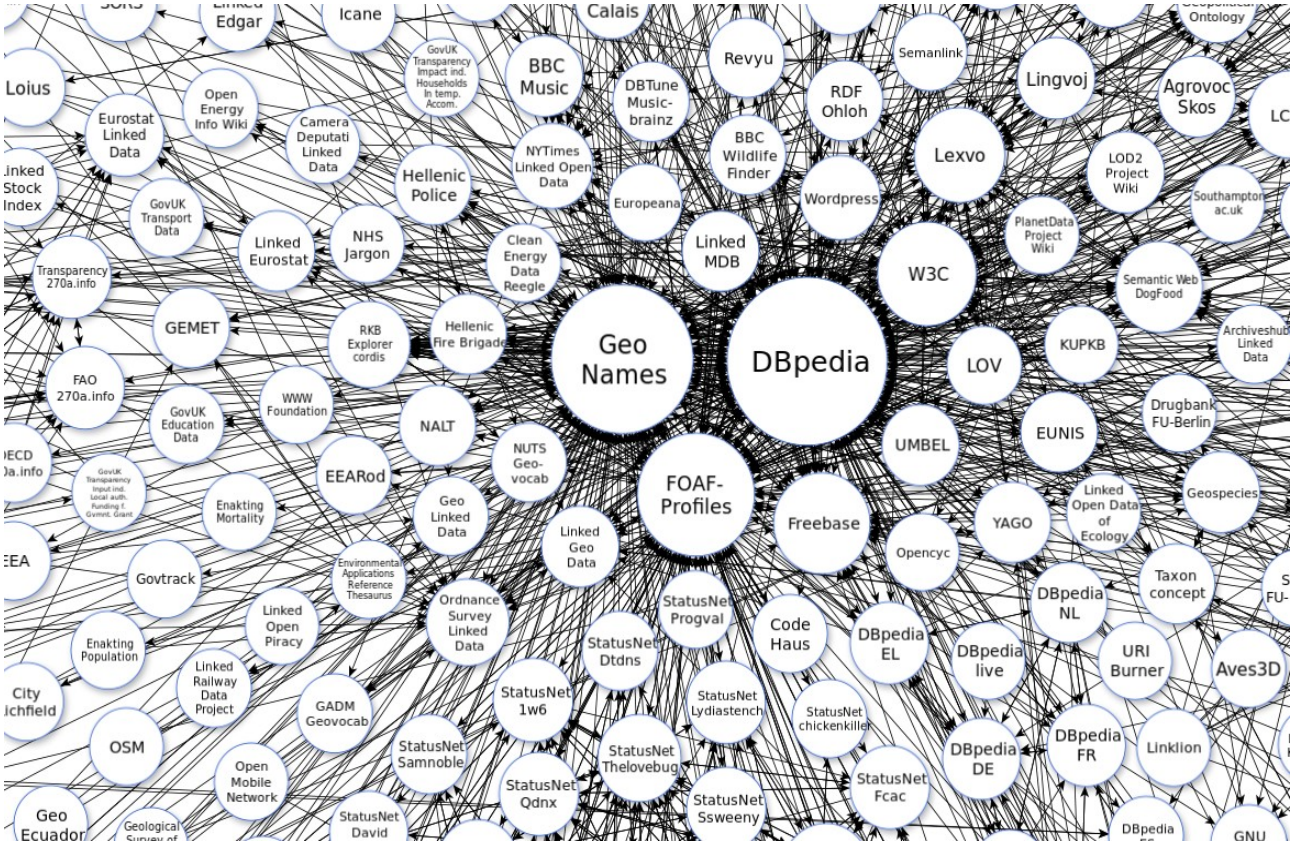

Abbildung 3: Linked Open Data on the Web (kleiner Ausschnitt) Quelle: linkeddata.org

Man sieht hier vielfältig semantisch vernetzte Knoten – prominent z. B. *Dbpedia*, aus *Wikipedia* extrahierte formale Daten. Aber auch viele andere bekannte Informationsquellen bieten in diesem Netz strukturierte Daten an und vernetzen sich mit anderen.

Bei näherer Betrachtung zeigt *Linked Open Data* auch Schwächen. Vieles ist experimentell, Links tauchen auf und verschwinden, Datenbestände sind teilweise schlecht gepflegt. Aber wird auch eine Menge Energie darauf verwandt, große Bestände wie *Wikipedia, OSM, CIA World Factbook*, für das *Semantic Web* zu erschließen. Es gibt auch einige Anwendungen, bei denen oft nicht bekannt ist, dass sie auf *Linked Open Data* beruhen. Berners-Lees Agenten sind jedoch (noch) nicht dabei.

Was daraus wurde

Heute ist die Dynamik woanders. Dafür sind auch Entwicklungen der Grund, deren Auswirkungen Berners-Lee in seinem Szenario nicht berücksichtigt hatte. War die Situation Anfang dieses Jahrhunderts noch weitgehend dadurch geprägt, dass das Netz, das *WWW*, eine verteilte Quelle für Informationen und Wissen war, erlauben es heute die praktisch unbegrenzten Speichermöglichkeiten, beliebig skalierbare Computerleistung und superschnelle Netze (*Clouds*) riesige Datenmengen, zentral (besser: proprietär) zu speichern und gleichzeitig das *WWW* als Ergänzung zu nutzen, für Bestände, die man selbst nicht speichern oder pflegen will.

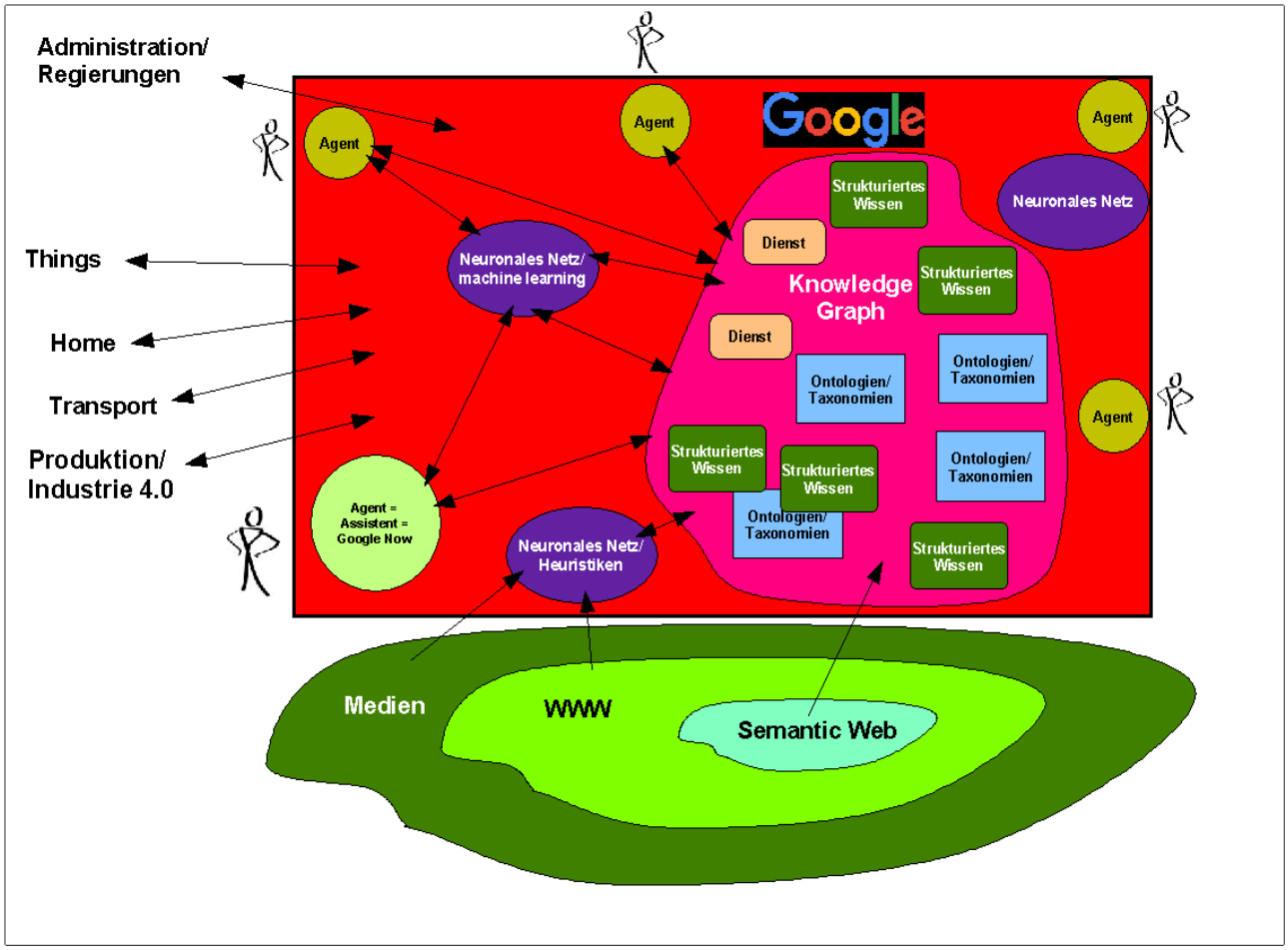

Abbildung 4: Infrastruktur bei Google

Die Strukturen, die Konzerne wie *Google, Apple, Microsoft, Amazon* mit hohem Aufwand und großer Energie aufbauen und in immer mehr Bereichen nutzen (Suchmaschinen, Agenten/Assisten-

ten, Marketing …) zeichnen sich zum einen durch eine höhere Anwendungsorientierung als zum Beispiel *Linked Open Data* aus. Sie benutzen aber für die Repräsentation von Wissen die gleichen oder verwandte Technologien. Ergänzt werden sie um mächtige Verfahren zur – weitgehend automatischen oder halbautomatischen – Erschließung und Strukturierung der Daten. Schematisch stellt sich das, was dort aufgebaut wird, etwa so wie in *Abbildung 4* dar. Hier habe ich *Google* als Beispiel genommen. Im Zentrum liegt der berühmte *Knowledge Graph*, ein riesiges internes semantisches Netz, das die Anwendungen von Google und ggf. von Partnern unterstützt. Google sieht seinen *Knowledge Graphen* so: *„The Knowledge Graph is Google's system for organizing information about millions of well-known "entities": people, places, and organizations in the real world.“ Das* gilt so auch für das Semantic Web. Weiter heißt es:*„Google's algorithms merge information about entities from many data sources. For some types of information, though, the best source of data is the entity itself.“* Diese Wissensbasis greift durchaus auch auf Wissen zu, das außerhalb etwa in Datenbanken im *Semantic Web* gespeichert ist. Am bekanntesten ist sicher *Wikidata*, dass Daten aus und für Wikipedia sammelt und von *Communities* unterstützt wird. Suchmaschinenoptimierer empfehlen inzwischen oft, mit dem eigenen Unternehmen auf *Wikidata* präsent zu sein, da *Google Wikidata* auswertet.

Zur Datenbasis kommen weitere wichtige Komponenten hinzu. Da ist zunächst die Datenerschließung:

- die Strukturierung nicht Strukturierten Wissens (z.B. aus dem *WWW*),
- die Auswertung von Suchanfragen (*the best source of data is the entity itself!*),
- die Auswertung anderer Anwendungen und Quellen (*Social Networks, News, Maps, Smart Phones* …, neuerdings *Home Automation,* Autos, *Internet of Things* …),
- die Erschließung multimedialer Inhalte (Bild – u.a. Gesichtserkennung, Biometrie, Video, gesprochene Sprache, Musik).

Das Phantastische daran ist, dass es praktisch keine – technischen – Grenzen für die semantische Vernetzung dieses Wissens gibt. Nicht alles ist zunächst sinnvoll, aber je dichter das Netz wird, um so vielfältiger sind die Anwendungsmöglichkeiten. (Berners-Lee bewirbt vernetzte Daten, mit dem Satz *„data can be used in unexpected ways“* und sieht darin einen fast grenzenlosen Raum für Kreativität – einige Vorkommnisse in jüngerer Zeit nähren allerdings Zweifel, ob das immer zum Wohle der Menschheit ist.)

Mustererkennung und neuronale Netze

Die Umwandlung von nicht strukturiertem Wissen in strukturiertes Wissen übernehmen vielfach neuronale Netze. Das sind Computer, deren Architektur einem einfachen Modell des menschlichen Gehirn nachempfunden sind. Diese Computer entwickeln erstaunliche Fähigkeiten, wenn es darum geht, Muster zu erkennen und zu analysieren. Muster, die natürlicher Sprache, gesprochener Sprache, Bildern, Filmen, Musik … Bedeutung geben. Vereinfacht gesagt, sind neuronale Netze in der Lage zu lernen, Mustern strukturierte Daten zuzuordnen.

Ein spektakuläres Beispiel, was ein großer Pool an strukturiertem Wissen und eine gute Erkennung von Sprache leisten können lieferte IBM's *AI*-Projekt *Watson* 2011, als *Watson*, ein Computer mit künstlicher Intelligenz, die amerikanischen Meister im *Jeopardy*-Quiz schlug.

Zum einen liefert unstrukturiert gespeichertes Wissen in *WWW*, *Wikipedia*, *Social Networks* … Input für neuronale Netze und andere Verfahren der künstlichen Intelligenz – zur strukturierten Er-

schließung. Andererseits generieren auch die meisten Anwendungen (neudeutsch Apps), die dieses Wissen nutzen, gleichzeitig neue Daten, also Lehrstoff für die Wissensnetze.

Oft wird für den Aufbau der semantischen Netze aber auch schlicht die Handarbeit von Vielen eingesetzt. Die beiden prominentesten Projekte um Wikipedia – Wikidata und Dbpedia – gehen hier unterschiedliche Wege. Während Dbpedia Daten automatisiert aus Wikipedia gewinnt, erfasst die Wikidata community Daten weitgehend manuell und macht sie u.a. auch einer Nutzung durch Wikipedia (sic!) zugänglich.

Rückkopplung

Wie oben erwähnt, liefern Anwendungen vielfach gleich die Daten, durch die sie selbst besser werden. Eine Suchmaschine kann aus dem Wissen, bei welchen Fragen welche Antworten akzeptiert werden, schließen, was generell oder individuell eine gute Antwort auf eine Frage ist. Bei selbst fahrenden Autos ist die Rückkopplungsschleife besonders evident: die (unstrukturierten) Daten, die Sensoren und Kameras einsammeln, müssen kurzfristig ausgewertet werden, und unter Einbeziehung langfristiger Daten (StVO, Karten, Infos über das Fahrziel) ausgewertet und angewandt werden.

Raum für Fehler

Abbildung 5: Strange Google Direct Answer (Amy Gesenhues on http://searchengineland.com)

Diese Technologien haben das Universum binärer Entscheidbarkeit endgültig verlassen. Das *WWW* ist voll widersprüchlicher Informationen. Wenn diese strukturiert erschlossen werden, ist auch das resultierende Wissen widersprüchlich. Damit müssen Anwendungen umgehen können. Als lernende Systeme müssen sie Wahrscheinlichkeiten bestimmen, was „falsch“ und „richtig“ ist. Das semantischen Netze vielfach noch am Anfang ihrer Lernzeit stehen – zeigen viele lustige Fehlleistungen von *Siri* und *Google Now* oder auch der intelligenten *Google* Suche (*Google Direct Answers).*

Auch von einem noch so guten, selbst fahrenden Auto ist kein fehlerfreies Fahren zu erwarten. Ein solches Auto ist aber einsetzbar, wenn es deutlich besser fährt als ein Mensch.

Und wir?

Die (für meine 15 Minuten) etwas ausführliche Schilderung der aktuellen Techniken soll zeigen, dass Wissensnetze – so wie sie heute bereits existieren, einen Umbruch bringen, gegen den Internet, Web, Mobiltelefonie und Smartphones laue Lüftchen waren:

- Problemorientierte, isolierte Datenbanken verlieren an Bedeutung. Die Zukunft gehört vernetzten Daten, deren Anwendungsbezug nicht von vornherein feststeht.
- Die Datenberge – und damit die Datennetze – wachsen derzeit rasant. Zwar findet auch ein Wachstum in der Fläche statt, aber die für ein breites Anwendungsspektrum am besten nutzbaren Daten sammeln sich bei Konzernen wie *Google, Apple, Microsoft* …
- Die sich abzeichnenden Anwendungsmöglichkeiten versprechen eine traumhafte Bereicherung der Möglichkeiten des privaten wie kommerziellen Lebens … wir werden für uns interessante Dinge erfahren, von denen wir sonst keine Ahnung hätten, die meiste Organisationsarbeit in unserem Alltag wird uns abgenommen, auf unsere Gesundheit wird besser aufgepasst, wir werden mehr Spaß haben, interessantere Leute kennenlernen, Demokratie wird einfacher.
- Der Traum von Privatheit ist – zumindest wie derzeit von Datenschützern proklamiert und gefordert – ausgeträumt. Assistenten die unser Leben begleiten, sind halt dann am nützlichsten, wenn sie viel über uns wissen. Selbst wenn Techniken zum Einsatz kämen, die dem Endanwender die Kontrolle seiner Daten ermöglichen – in der Praxis ist er nicht in der Lage, seine Daten auch nur zu überschauen. Der Berg Papier, den der Österreicher Max Schrems als „seine *Facebook*-Daten“ bekam, ist nur ein lächerlich kleiner Teil dessen, was in die semantischen Netze eingeht. Harte Zeiten für Datenschützer und harte Zeiten für die Gesetzgebung.
- Auf den Hütern der Daten, die sich oft auch weitgehende Rechte an diesen einräumen lassen, lastet eine enorme Verantwortung. „Wir sind die Guten“ sagt *Google*. Aber *Google* und die anderen Firmen in diesem Geschäft sind nicht neutrale, philanthropische Organisationen sondern bei Firmen, die Geld verdienen müssen und wollen – und deren Kapital eben diese Daten sind. Allerdings sind sie auch auf das Vertrauen ihrer Kunden angewiesen, dass sie sich auf lange Sicht erhalten müssen.
- Sicherheit der Daten? Es gibt keine 100%ige Sicherheit. Man sollt mal über Open Source nachdenken (Verbesserung der Kontrollmöglichkeiten), mehr Kryptographie im Netz (wie es die IETF derzeit fordert), … das Alles verbessert die Sicherheit, ist aber nicht die Lösung.
- Auswirkungen auf Jobs? Selbstverständlich. Neue Jobs in Datenaufbereitung, Verifikation von Daten, Qualitätskontrolle für Anwendungen … werden entstehen. Viele Verwaltungsjobs (Sekretariate, Assistenten) werden sich neu ausrichten müssen oder verschwinden.

Da kommt Einiges auf uns zu. Zunächst einmal eine hoffentlich spannende Diskussion.

Referenzen (Auswahl):

Berners-Lees *Semantic Web* Artikel:

http://www-sop.inria.fr/acacia/cours/essi2006/Scientific%20American_%20Feature%20Article_%20The%20Semantic%20Web_%20May%202001.pdf

***Semantic Web* bei *W3C*:**

http://www.w3.org/2013/data/

Zu *Googles Knowledge Graph*:

https://googleblog.blogspot.de/2012/05/introducing-knowledge-graph-things-not.html

http://www.zdnet.com/article/google-knowledge-graph-this-is-why-they-changed-their-privacy-policy/

Zu *Googles Knowledge Graph, Wikidata* und *SEO*:

http://www.searchenginejournal.com/wikidata-meets-google-knowledge-graph/130459/

***When Google Gets It Wrong: Direct Answers With Debatable, Incorrect & Weird Content* (Alaaf: Tribut an den heutigen 11.11. in Köln):**

http://searchengineland.com/when-google-gets-it-wrong-direct-answers-with-debatable-incorrect-weird-content-223073

Neuronale Netze:

http://pages.cs.wisc.edu/~bolo/shipyard/neural/local.html

2 *AI* Pioniere, die im Statement etwas zu kurz kamen:

***IBMs Watson* Projekt**:

http://research.ibm.com/cognitive-computing/watson/

Wolfram Alpha:

http://www.wolframalpha.com/