Viele Daten – Viel Wissen?

Dr. Matthias Nagel, Simba n³

Sehr verehre Damen und Herren,
ich werde oft gebeten, immer über so Allerwelts-Zeugs wie IoT, Industrie 4.0, Digitalisierung, Big Data, Business Intelligenz zu reden.

Und nun soll ich heute auf dem Mind Slam über ganz konkrete Dinge wie **Daten** und **Wissen** sprechen - und kann für einen Slam nicht mal reimen!

Der Titel **„Viel Daten - Viel Wissen Fragezeichen“** hat es erst auf den 2. Blick in sich!

Was sind denn **viele Daten**? Und **was ist viel Wissen**?

Haben die Veranstalter damit vielleicht das Wissen von ausgewiesenen Spezialisten wie z.B. über die Aufstellung der Fußballnationalmannschaft 1954 zur Weltmeisterschaft gemeint? Mit solchem Wissen kann man manchen Ortes sogar viel Geld machen.

Ich glaube das eher nicht. Das Thema ist also doppelbödig, deshalb muss ich mich an den einzelnen Worten des Titels entlang hangeln, um zu versuchen, die semantischen Kategorien „Daten“ und „Wissen“ irgendwie zusammen zu bringen.

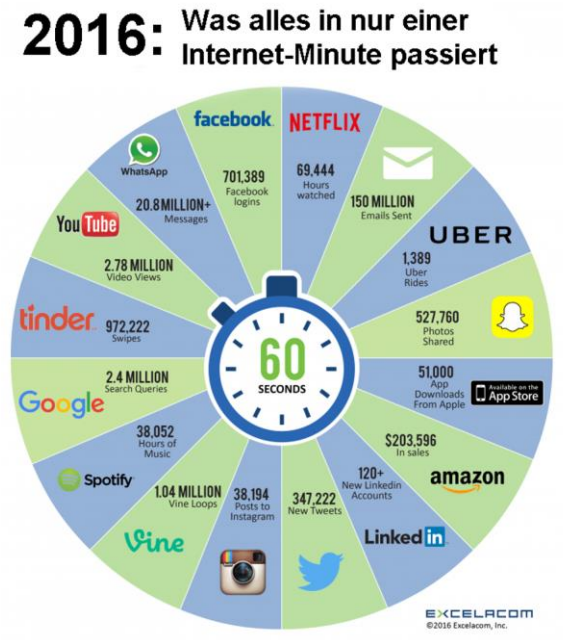
Was sind also **viele Daten?**

Das, was in einer Internet-Minute passiert ist wirklich echt viel. Aber da gibt es ja noch all das, was jetzt schon an Daten irgendwo digital oder analog rumliegt.

WAS SIND VIELE DATEN?

90 Prozent aller gespeicherten Daten wurde in letzten zwei Jahren erzeugt.

Simba n³ SOFTWARE INNOVATIVE LÖSUNGEN



SIMBA N³ | OKTOBER 2017 | CHEMNITZER MIND SLAM 2

Oder sind Daten aus einem Smart Home wie diese hier, aufgenommen am 9. und 10. Oktober in der Wohnung von Freunden in Hamburg, in ihre Summe Daten mit viel Wissen?

Man kann aus Raumtemperatur und Luftfeuchtigkeit nämlich auf die Zahl der Personen schließen, die sich in der Wohnung aufhalten, wann sie die Wohnung verlassen oder wieder zurückgekommen. Hat man dazu den Stromverbrauch im Sekundenbereich im Datenzugriff, kann man das Muster der Geräte erkennen, und aus den Hell- Dunkel-Mustern des Fernsehers kann man auf den Film schließen, der darauf angesehen wird.

Nun, dass wir an dem Abend gemeinsam das Nick Cave Konzert in Hamburg besucht haben, wird daraus nicht klar. Dazu müsste irgendwer noch die E-Mails von Gast und Gastgeber durchsuchen oder die Daten des Navis meines Autos tracken.

Sie glauben mir das alles nicht? Wir können uns dazu gern in der Pause unterhalten!

Lesson learned: Wir hinterlassen überall elektronische Spuren - oft unabhängig davon ob wir das wollen oder nicht. Dies Spuren können für einige interessant sein - von denen wir auf keinen Fall wünschen, dass sie über uns alles wissen.

Auf jeden Fall sind die Spuren und Daten für das Amazon‘ s dieser Welt echt Gold wert.

Per se wissen die Schlapphüte erstmal noch nicht allzu viel. Also haben sie Selektoren und nutzen Metadaten der Kommunikation – **Wer mit Wem? Wann? Wo?** - um damit alle Daten, die sie im Zugriff haben, auszuwerten. Haben sie etwas gefunden, was zu den Selektoren passt -Zack - läuft eine Maschinerie an. Ihr sorgloses „Ich habe ja nichts zu verbergen!“ hilft da erst mal nicht viel weiter.

Ist für Wissen die Menge an Daten wichtig?

IST DIE MENGE AN DATEN WICHTIG?

Nicht unbedingt. Aber man braucht zu Daten Metainformationen.

Simba n³ SOFTWARE INNOVATIVE LÖSUNGEN

	Akuter Stress	Chronischer Stress	**Burnout**	Optimalwert
DHEA	452.3	853.2	**123.3**	300-600
Cortisol (7.oo)	6.2	2.1	**1.5**	4-12 (7.oo)
(12.00) ng/ml	3.2	1.5	0.9	3-6 (12.00)
(13.oo)	1.9	1.8	0.8	2-5 (13.oo)
(20.oo)	0.9	1.0	0.5	<1.5 (20.oo)
Epi	29.4	1.3	1.8	8-12
Noradrenalin	96.5	94.2	**22.3**	30-55
Dopamin	130.6	255.8	**99.4**	125-175
Serotonin	162.0	52.8	**67.8**	175-225
GABA	22.4	7.3	9.2	1.5-4.0
Glutamate	13.5	56.2	63.1	10-25
PEA	300.0	734.2	324.5	175-350
Histamine	28.0	18.2	9.5	10-25

Quelle: www.antox.de

SIMBA N³ | OKTOBER 2017 | CHEMNITZER MIND SLAM 3

Offenbar nicht zwingend! Für die **Diagnose eines Burnouts** reichen offenbar schon 5 Werte aus. Aber zuvor muss der Arzt die initiale Idee haben, den Patienten auf Burnout hin zu untersuchen.

Lesson learned: Voraussetzung für viele Entscheidungen sind relevante Daten und Metawissen für das Problem.

Die große Schaar der Beamten des sympathischen Herrn auf der obigen Bild sammeln viele Finanzdaten von uns allen ein und sind geschult darin, möglichen Tricksereien zu finden. Mit den Daten der folgenden Liste mit nur wenigen Spalten werden sie vermutlich aber nur wenig anfangen können:

3567283200	G10006	4	3563740800	100822235
3553977600	G10035	7	3552249600	111772535
3559334400	G10013	5	3558297600	102546635
3566073600	G10006	8	3565382400	109045635
3552595200	G10013	9	3551472000	101871935
3564864000	G10006	1	3561062400	119575335
3577392000	G10016	1	3576096000	100822236
3553632000	G10006	4	3552422400	100822237
3552076800	G10007	3	3550003200	110800235
3531859200	G10009	1	3529008000	100822238
3539635200	G10024	7	3535920000	101871936
3579897600	G10048	0	3577737600	115873435
3561148800	G10013	0	3557174400	119575336
3535056000	G10048	4	3533155200	110492235
3554755200	G10006	7	3551385600	100822239
3557088000	G10008	0	3556137600	100822240
3533328000	G10009	5	3531686400	119575337
3539808000	G10020	9	3538425600	100822241
3570652800	G10009	0	3566764800	100822242
3532118400	G10037	9	3531254400	115873436
3535401600	G10024	8	3533587200	101871937
3561321600	G10006	8	3560371200	100822243
3555446400	G10021	9	3553286400	102546636
3541795200	G10032	6	3538339200	115873437
3548448000	G10024	9	3545510400	111772536
3545078400	G10009	2	3543955200	100822244
3569702400	G10013	5	3565468800	100822245
3573158400	G10020	8	3571516800	100822246
3556915200	G10016	1	3555100800	119575338
3560371200	G10018	8	3559593600	100822247
3542745600	G10018	5	3541017600	100822248
3559593600	G10018	3	3558643200	108463435
3555878400	G10016	4	3554496000	100822249
3531686400	G10018	0	3530649600	100822250
3544819200	G10013	0	3542054400	110800236
3578083200	G10021	2	3575750400	100822251
3544560000	G10018	8	3542745600	115873438

Es fehlen in der Tabelle nämlich die Metainformationen, um was es sich dabei handeln könnte. Auch die Selektoren der Schlapphüte gehen damit ins Leere.

Allein mit den Überschriften wird daraus eine wahre Schatztruhe für jedes Unternehmen!

DATEN ZU EREIGNISZEITEN SIND SEHR NÜTZLICH!

Und solche Daten gibt es in jedem Unternehmen.

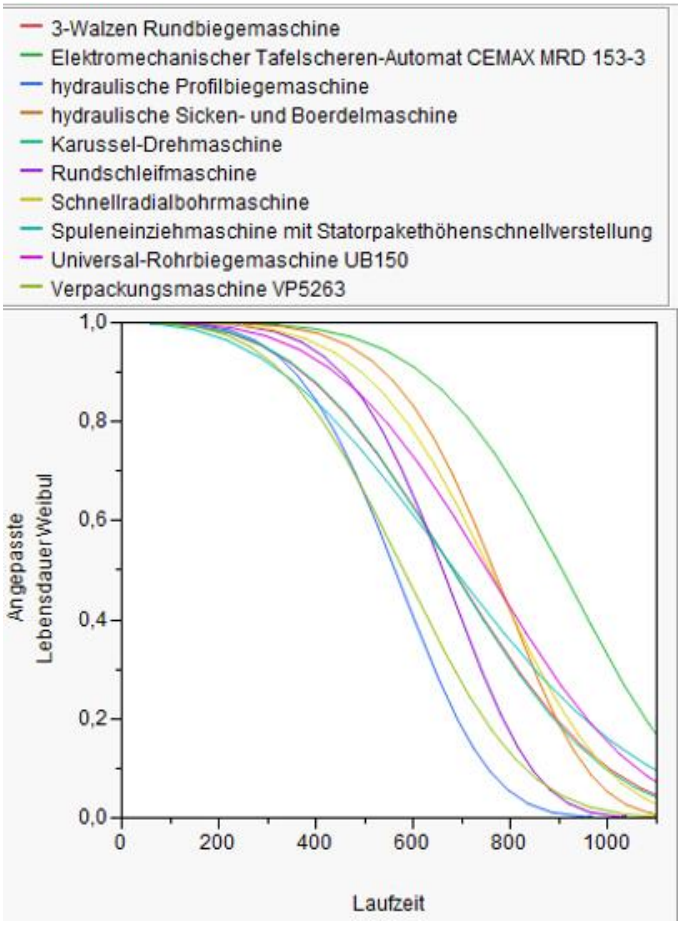
Simba n³ SOFTWARE INNOVATIVE LÖSUNGEN

AusfallvonDatumText	ChargennummerTeilText	FehlerursacheValue	InbetriebnahmedatumValue	SeriennummerText
15.01.2017	G10006	4	05.12.2016	100822235
14.08.2016	G10035	7	25.07.2016	111772535
15.10.2016	G10013	5	03.10.2016	102546635
01.01.2017	G10006	8	24.12.2016	109045635
29.07.2016	G10013	9	16.07.2016	101871935
18.12.2016	G10006	1	04.11.2016	119575335
12.05.2017	G10016	1	27.04.2017	100822236
10.08.2016	G10006	4	27.07.2016	100822237
23.07.2016	G10007	3	29.06.2016	110800235
02.12.2015	G10009	1	30.10.2015	100822238
01.03.2016	G10024	7	18.01.2016	101871936
10.06.2017	G10048	0	16.05.2017	115873435
05.11.2016	G10013	0	20.09.2016	119575336
08.01.2016	G10048	4	17.12.2015	110492235
23.08.2016	G10006	7	15.07.2016	100822239
19.09.2016	G10008	0	08.09.2016	100822240
19.12.2015	G10009	5	30.11.2015	119575337
03.03.2016	G10020	9	16.02.2016	100822241
23.02.2017	G10009	0	09.01.2017	100822242
05.12.2015	G10037	9	25.11.2015	115873436
12.01.2016	G10024	8	22.12.2015	101871937
07.11.2016	G10006	8	27.10.2016	100822243
31.08.2016	G10021	9	06.08.2016	102546636
26.03.2016	G10032	6	15.02.2016	115873437
11.06.2016	G10024	9	08.05.2016	111772536
03.05.2016	G10009	2	20.04.2016	100822244
12.02.2017	G10013	5	25.12.2016	100822245
24.03.2017	G10020	8	05.03.2017	100822246
17.09.2016	G10016	1	27.08.2016	119575338
27.10.2016	G10018	8	18.10.2016	100822247
06.04.2016	G10018	5	17.03.2016	100822248
18.10.2016	G10018	3	07.10.2016	108463435
05.09.2016	G10016	4	20.08.2016	100822249
30.11.2015	G10018	0	18.11.2015	100822250
30.04.2016	G10013	0	29.03.2016	110800236
20.05.2017	G10021	2	23.04.2017	100822251
27.04.2016	G10018	8	06.04.2016	115873438

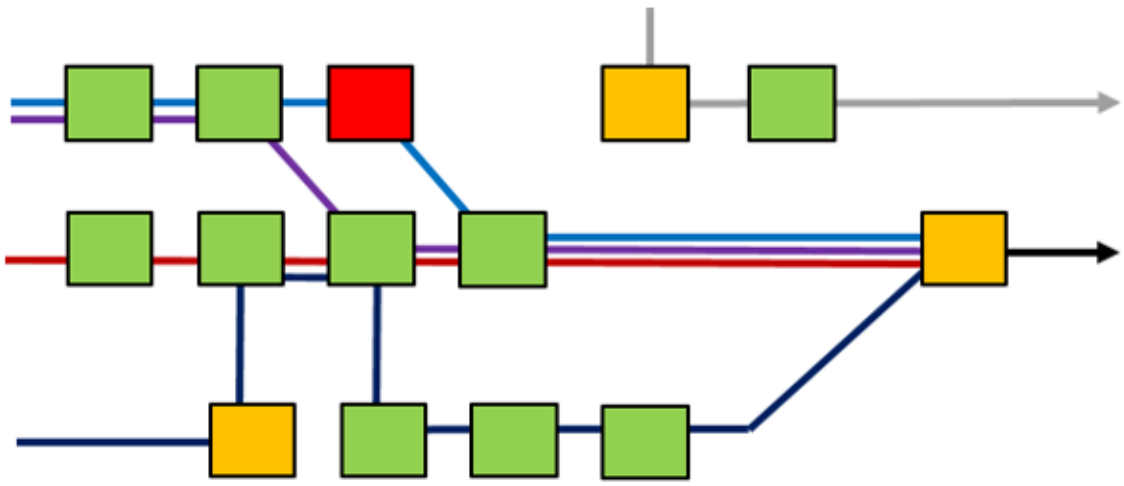
In jedem Unternehmen kann man nämlich die Zeiten bis zu einem Ereignis und die Ursache dafür recht leicht aus den Prozessen erfassen. Also die Zeiten, bis ein Prozess gestört wird, und danach die Zeit bis zur erneuten Funktionsfähigkeit des Prozesses. Und man braucht dafür nicht einmal unbedingt Sensoren.

Wir können z.B. aus solcherart Daten für jede Maschine und jede Ursache für deren Stillstand ein Modell für die Wahrscheinlichkeiten eines solchen Ausfalls berechnen.

In der nächsten Abbildung wurden für obigen Daten Maschinen zugeordnet und für deren Ausfall Modelle berechnet. Wir wissen durch die Modelle, nach welcher Zeit oder welchem Verbrauch die Maschinen wahrscheinlich ausfallen werden.



Das ist noch lange nicht alles. Man kann damit die restliche Laufzeit jeder Maschine in einem bestimmten Zeitraum ermitteln, um dafür Ersatzteile bereitzustellen – nicht zu viele, aber auch nicht zu wenige. Man kann damit die Ausfallkosten zu planen oder sogar die Funktionsfähigkeit ganzer komplexer Produktionssysteme bewerten und beurteilen.



Die rot markierte Maschine wird in einem nahen Zeitintervall mit hoher Wahrscheinlichkeit ausfallen. Sie wird durch den Ausfall die Produktionsstecke in der Mitte des Schemas massiv stören. Hier ist also schnelles Handeln angesagt.

Die gelben Maschinen sollten ebenfalls zeitnah proaktiv gewartet werden - z.B. dann, wenn gerade der Servicetechniker im Haus ist.

Die Modelle sind außerdem sogar auf alte Maschinenbestände anwendbar.

Solche Wahrscheinlichkeitsmodelle befeuern das **Hype Thema Predictive Maintenance** und Industrie 4.0.

Wer sich für das Thema näher interessiert, sei auf den Vortrag von Prof. Riedel und mir auf dem Kongress **TBI2017 am 10.11.2017 an der TU Chemnitz** verwiesen.

Lesson learned: Data Scientist können mit Daten und geeigneten Metainformation mächtige Algorithmen bereitstellen, die Ingenieuren bei ihren Entscheidungen unterstützen, z.B. um Maschinen proaktiv zu warten oder die Gesamtanlageneffektivität (OEE) zu ermitteln, um damit Prozesse optimieren zu können.

Lesson learned: Rohe Maschinendaten kann man - wenn sie dazu noch verschlüsselt sind und Schlüssel und Metadaten ausschließlich im Unternehmen bleiben - sicher in der Cloud vorhalten.

Lesson learned: Die Menge an Daten allein ist für Wissen nicht entscheidend. Man benötigt immer zusätzliche Informationen zu Daten. Die Geheimdienste brauchen ihre Sektoren, wir Zivilisten benötigen zumindest den Sinn und die Zusammenhänge, warum die Daten erfasst wurden.

Daten können richtig nützlich sein!

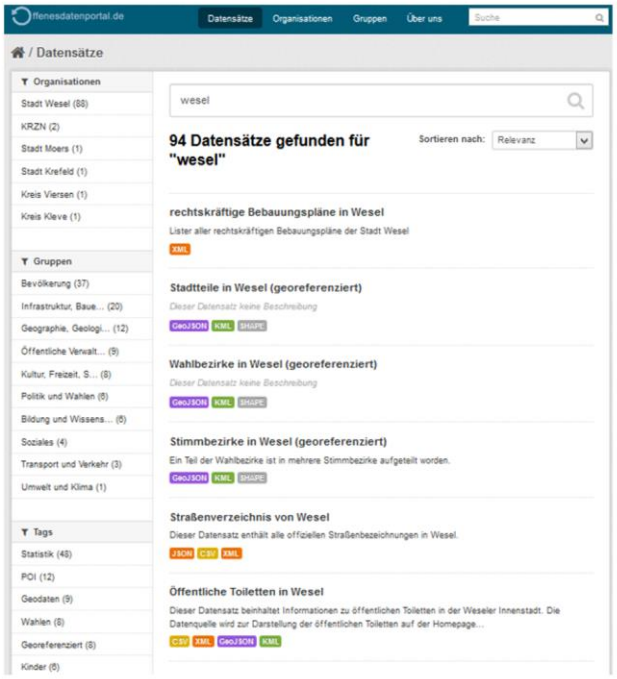
Vermutlich genau deswegen hat im Mai „… der Bundestag ein halbgares Open-Data-Gesetz beschlossen“, „…mit den digitalen Behördendaten maschinenlesbar und entgeltfrei öffentlich zugänglich gemacht werden sollen. Einen Anspruch darauf gibt es aber nicht, die Ausnahmen sind groß.“ (Zitat Heise online).

Hier ein Beispiel eines von vielen Open Data Portalen, zu dessen Inhalten allein die Stadt Wesel besonders fleißig 94! Datenbestände beigesteuert hat. Wir finden da u.a. Wahlbezirke und das Straßenverzeichnis – vermutlich als jetzt amtliche Ergänzung zu Google Maps gedacht und als besonderen Knüller die 5 öffentlichen Toiletten mit ihren Geo-Koordinaten - aber ohne die vielleicht auch relevante Information, ob dort aktuell auch Papier vorhanden ist.

OPEN DATA…

… sollten nützlich sein!

Simba n³ SOFTWARE INNOVATIVE LÖSUNGEN



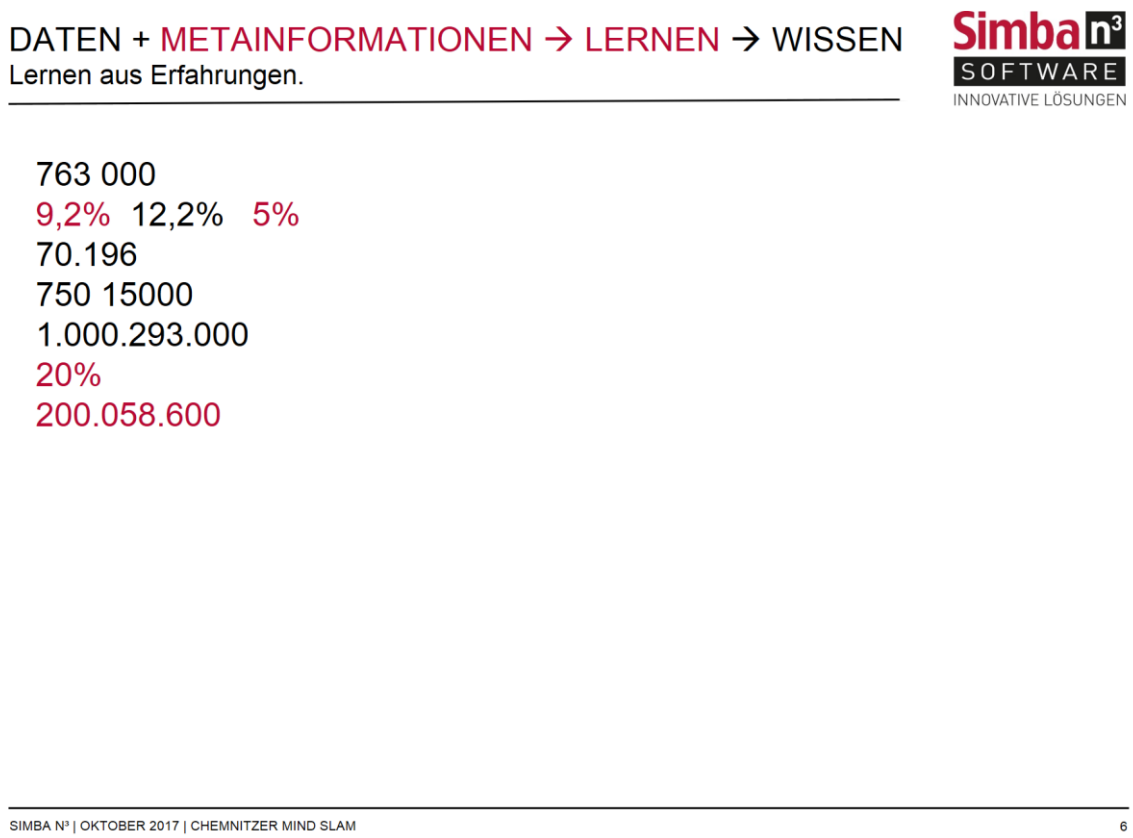
SIMBA N³ | OKTOBER 2017 | CHEMNITZER MIND SLAM 5

Vom Amtsschimmel in Wesel wurde Open Date offensichtlich etwas falsch verstanden.
Das es aber auch anders geht, zeigt das Open Data Portal der Bahn. **Unbedingt ansehen!**

Lesson learned: Open Data sollten für **Dritte Nutzer immer auch einen Nutzen** haben.

Lernen aus Daten

Lassen Sie mich jetzt noch ein Beispiel bringen, um den Titel des Vortrags um die Begriffskette **Lernen aus Daten** anzureichern.



Sie werden es vielleicht nicht sofort glauben, aber obige Zahlen sind für die Zukunft unseres Landes sehr wichtig.

Denn: **763 000** ist in etwa die Anzahl der jährlichen Geburten in Deutschland.

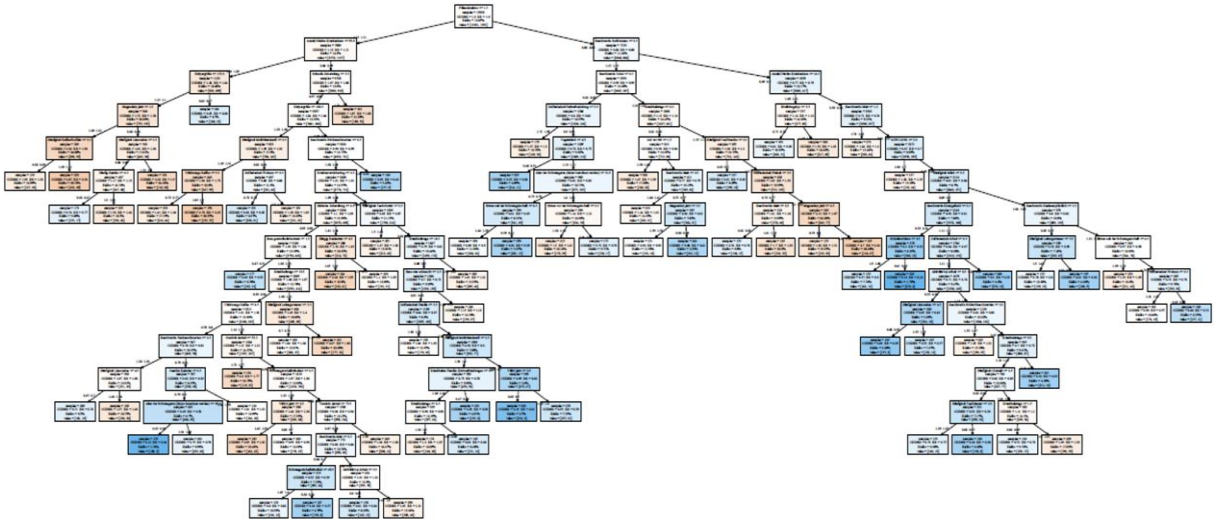
Rot hervorgehoben: **9,2%** ist die Periodenzahl der Frühgeborenen in Deutschland, **5%** die in den skandinavischen Ländern. Das reiche Deutschland liegt fast doppelt so hoch. Noch höher ist die Rate in den USA: 12,2% - vermutlich wegen der übergewichtigen, älteren und sich falsch ernährenden Schwangeren. **Halt:** Das war eine **reine Vermutung von mir**, ich ziehe das als **Fake News** zurück!

Wir haben also je Jahr in Deutschland über **70 000** Frühgeborene. Im Unterschied zu einem gesunden Baby, für das bei einer Geburt etwa 750 € entstehen, fallen für den zu frühen kleinen Menschen durchschnittlich 15 **000 €** an, bis er aus der Klinik in der Obhut der Eltern übergeben werden kann. Frühgeborene als die größte Gruppe kranker Kinder belasten unser Gesundheitssystem also mit ca. **1 Mrd. € pro Jahr.**

Jetzt die gute Nachricht: Man kann durch gesunde Lebensweise, gesundes Essen und gewisse Vorsorgemaßnahmen während der Schwangerschaft den Prozentsatz der Frühgeborenen nachweislich um ca. 20% senken! Das haben die Forscher von FBE, mit denen ich seit Jahrzehnten zusammenarbeite, aus Fragebögen von 60 000 Schwangeren und aus mehr als 20 000 Schwangerschaftsverläufen mit ärztlicher Befundung ermittelt.

20 % entsprechen immerhin ca. 200 Mio. pro Jahr, die man gut für Präventionen einsetzen **könnte.**

FBE wollten nun von uns wissen, welche Kombinationen aus über 500 Faktoren für das Risiko einer Frühgeburt relevant sind - und in welchem Maße. Wir als Data Scientists haben dazu eine Machine Learning-App programmiert, bei der man beim Durchlaufen eines Modell- Baumes bereits nach wenigen Fragen weiß: **Ob**, **Wie hoch** das Risiko einer Frühgeburt ist und **Durch welche Faktoren** es befördert wird.



Die schlechtere Nachricht: Unser Gesundheitssystem ist selbstverwaltet. Es gibt daher bis heute leider keine Ziffer, die Frauenärzte für eine zeitaufwändige Befragung zur Lebensweise der Schwangeren und für eine Frühgeborenen-Prävention abrechnen können. Daher findet das auch nicht statt.

Daher hatten FBE und wir die Idee, (zunächst) in den Wartezimmern großer Frauenkliniken Tabletts mit einer App auszulegen, in der das Risiko der werdenden Muttis durch gezielte Beantwortung von Fragen ermittelt wird. Selbstverständlich sind es **weder wir** noch **die App**, die der Schwangeren vermutlich sehr unsensibel mitteilt, dass ihr Risiko auf Frühgeburt dreimal so hoch als normal ist, weil sie z.B. zu dick ist, Kampfsport betreibt oder sich ungesund ernährt. Das übernimmt Arzt/Ärztin bei der Untersuchung. Auf einen Blick sieht man in der App das Ergebnis. Arzt/Ärztin können also digital unterstützt **ihr Fachwissen an die Schwangere in psychologisch geeigneter Form übermittelt.** Durch die App bekommt man mit den Risikofaktoren auch die Hinweise, was zu tun ist, um das Risiko für die restliche Zeit der Schwangerschaft zu verringern. Ohne zusätzlichen Zeitaufwand kann so der Schwangeren ein individuelles präventives Programm erstellt werden.

Das ist ein Beispiel dafür, wie man die wertvolle „Ressource Arzt“ zum Nutzen des Patienten digital optimal unterstützen kann, ohne ihn in seiner verantwortungsvollen Rolle am Patienten zu bevormunden.

Lesson learned: Machine Learning Verfahren sind überall in Produktion, Wirtschaft und Medizin gleichermaßen gut einsetzbar. Man benötigt dazu lediglich mögliche Einflussdaten auf Ereignisse - und **ausreichend viele** Daten zu Ereignissen. Data Scienstists sind zusammen mit den Fachleuten meist in der Lage, für das Problem geeigneten Verfahren und Algorithmen auszuwählen, um all das dann zu einer maßgeschneiderten Lösung umzusetzen.

Daten – Metainformationen – Lernen – Wissen - Vermittlung

Wir haben nun die Kette von Daten zu Wissen über **Metainformationen, Lernen und der Vermittlung des Wissens** geschlossen. Jetzt mach der Titel auch Sinn.

Ich möchte aber zum Schluss noch den Byung-Chul Han zu Wort kommen lassen.

Der Koreaner lebt und lehrt als Philosoph in Deutschland und veröffentlicht auch in deutsch. Alle seine durchweg schmalen Bändchen sind überaus voll an Ideen und sehr inspirierend zu lesen. So schreibt er z.B. über Big Data und Korrelation vs. Wissen Folgendes:

BYUNG-CHUL HAN & BIG DATA
Korrelationen vs. Wissen

Simba n³ SOFTWARE INNOVATIVE LÖSUNGEN

„Selbst die größte Ansammlung von Informationen, Big Data verfügt über sehr wenig Wissen. Anhand von Big Data werden Korrelationen ermittelt. Die Korrelation besagt: Wenn A stattfindet, so findet auch B statt. *Warum* es so ist, *weiß* man aber nicht.

Die Korrelation ist die primitivste Wissensform, die nicht einmal in der Lage ist, das Kausal-verhältnis, d.h. das Verhältnis von Ursache und Wirkung, zu vermitteln. *Es ist so*.
Die Frage nach dem Warum erübrigt sich hier.

Es wird also nichts *begriffen.* Wissen ist aber Begreifen.
So macht Big Data das Denken überflüssig.
Wir überlassen uns bedenkenlos dem Es-Ist-so.“

Für uns kann das nur heißen:
Bei Machine Learning nie das Denken vergessen.
Das macht Data Scientist aus.

SIMBA N³ | OKTOBER 2017 | CHEMNITZER MIND SLAM 8

Lesson learned: Wenn man also schon Maschinen und Algorithmen die Arbeit machen lässt, **darf der Mensch nie** aufhören, über die Ergebnisse und Folgen von deren Arbeit nachzudenken.

Sonst landet man sehr schnell bei Fake News und kommt vermutlich zu der Erkenntnis, dass das nachweisliche Absinken der Niederlande mit dem starken Bevölkerungswachstum dort oder – noch schlimmer! mit der Zuwanderung von Flüchtlingen zutun hat (Korrelation!)

Viele Dank für Ihr Interesse

Fazit:

Meine 12 Minuten des Mind Slams sind um, der Text ist etwas länger, als das, was ich gesagt habe.

Wenn Sie dennoch auf einer Zusammenfassung bestehen, kopieren Sie doch einfach die mit **Lesson learned** gekennzeichnete Abschnitte hierher. Damit ist zur Zusammenfassung alles gesagt.

Simba n³ SOFTWARE
INNOVATIVE LÖSUNGEN

MEHR INTERESSE AN DATA SCIENCE & ANALYTICS?

Dr. Matthias Nagel
Managing Director

Simba n³ Software GmbH
Dr.-Friedrichs-Straße 42
08606 Oelsnitz

Telefon: +49 (37421) 7224-0

Email: matthias.nagel@nhochdrei.de
Internet: www.nhochdrei.de

SIMBA N³ | OKTOBER 2017 | CHEMNITZER MIND SLAM