IIBA International Institute of Business Analysis™

Herausforderungen von Big Data

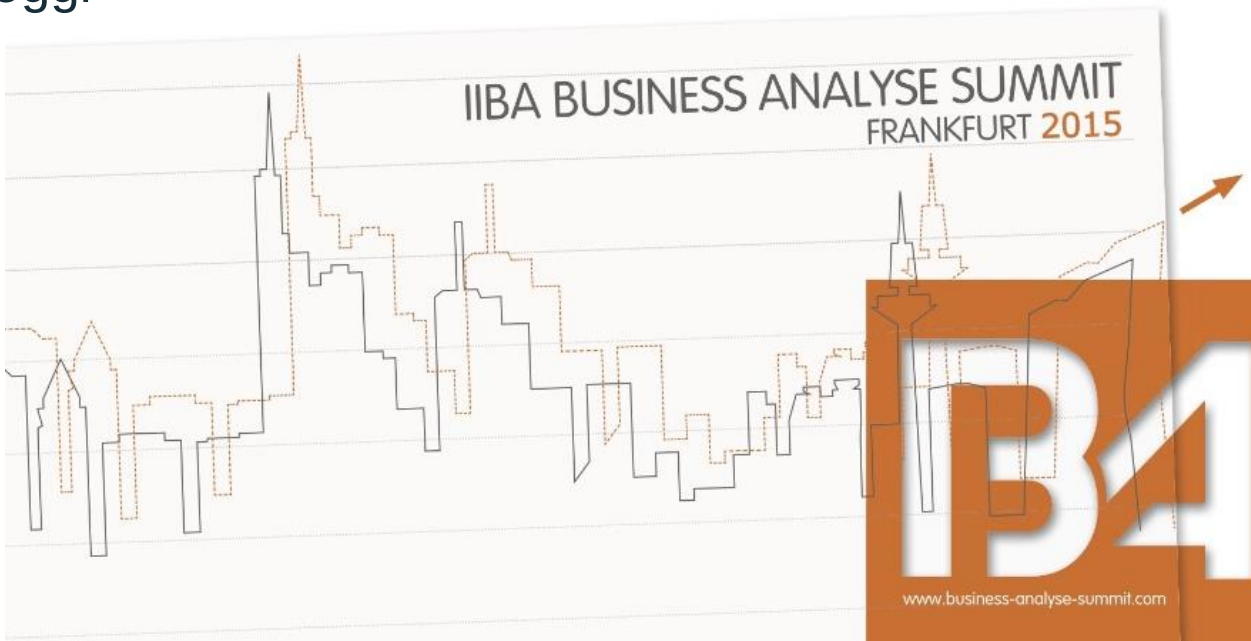
und ihr Impact auf die Business Analysis

Prof. Dr. Giampiero Beroggi

Frankfurt, 2015 Sep. 25

IIBA Austria Chapter

IIBA Germany Chapter



Was heisst «Big»?

Moore's Law

Volume

- Internet Archiv hat über 15 PB
- Archive: Facebook (40 PB), Yahoo! (60 PB), Ebay 40 (PB)

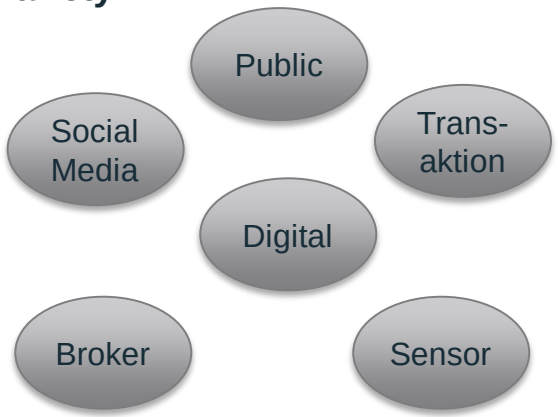
Wert	Abk.	Bezeichnung
1000	kB	kilobyte
1000^2	MB	megabyte
1000^3	GB	gigabyte
1000^4	TB	terabyte
1000^5	**PB**	**petabyte**
1000^6	EB	exabyte
1000^7	ZB	zettabyte
1000^8	YB	yottabyte

Velocity

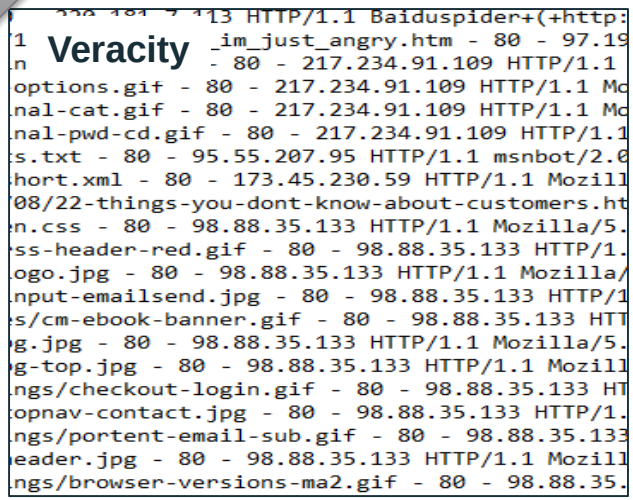
- Microsoft: Migration von Hotmail zu Outlook über 150 Petabytes in sechs Wochen migrierte
- IBM's neue Supercomputer «Summit» und «Sierra» werden ab 2017 mehr als 17 PB/Sek verschieben (= 100 Milliarden Facebook Photos)
- Tägliches Handling: Facebook (100 TB), Twitter (8 TB), Ebay (50 TB), AT&T (30 PT), Google (24 PB)
- Netflix liefert ca. 3 TB/Sek Videomaterial an seine Kunden

4V

Variety



Veracity



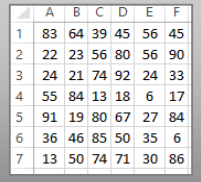
Big Data Life-Cycle

Business

Technology Push

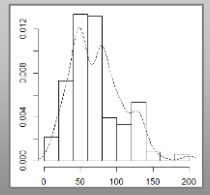
Technology Pull

Big Data

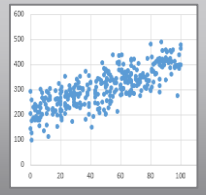


	A	B	C	D	E	F
1	83	64	39	45	56	45
2	22	23	56	80	56	90
3	24	21	74	92	24	33
4	55	84	13	18	6	17
5	91	19	80	67	27	84
6	36	46	85	50	35	6
7	13	50	74	71	30	86

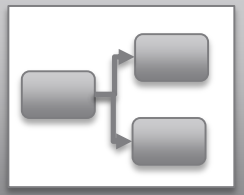
Smart Data



Information



Wissen



Entscheid

Informatik

Data Mining

Business Intelligence

Business Analysis

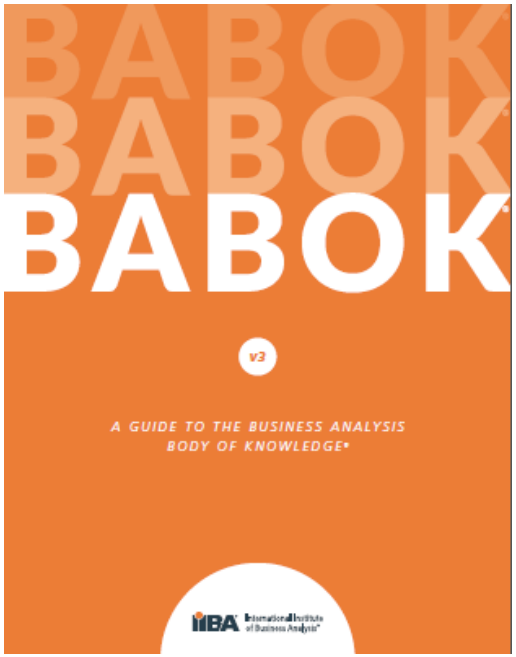
Menschliches Gehirn macht Big Data

- Pro Sekunde treffen auf den Menschen 50 GB ein.
- Davon nehmen wir lediglich 250 Bytes (0.25 KB) wahr, der Rest wird gefiltert.
- Somit nehmen wir nur ein Zweihundertmillionstel von dem wahr, was um uns passiert.
- Täglich erinnern wir uns an 34 GB Information.
- Das Gedächtnis kann gleichzeitig nur 6-7 Informationen speichern (z.B. Namen).
- Gehirn arbeitet mit nur 15 Watt Leistung, deshalb muss es effizient sein und Informationen filtern.

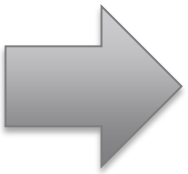
Big Data und Statistik

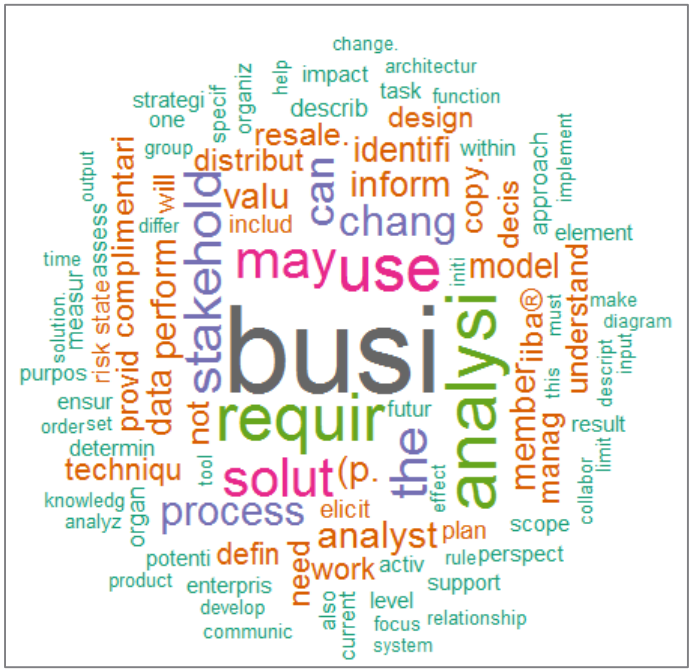
- «Big» aus Sicht der Statistik basiert auf vier Aspekte:
  1. **Zustand**: In welchem Zustand befinden sich die Daten für die Analyse; sind sie bereinigt (Ausreisser eliminiert, fehlende Daten ergänzt)?
  2. **Ort**: Wo befinden sich die Daten; auf unterschiedlichen Trägern, in unterschiedlicher Form (z.B. relationale DBs)?
  3. **Population**: Von wie viele Zielgruppen und Populationen kommen die Daten; sind es Primär- oder Sekundärdaten?
  4. **Umfang**: Besser repräsentative, hochqualitative und kleine Stichproben, als grössere und unzuverlässige Stichproben.
- Die Statistik sammelt Daten mit «Scheinwerfer-Methode» (auf Hypothesen basierend), während die IT Daten mit «Garbage-Can» Modell einsammelt (alles was digitalisiert werden kann, wird gesammelt).
- Für die Statistik kann (je nach Hypothese) eine Stichprobe von rund 20 Beobachtungen schon genügen.

Text-Analysen



2MB



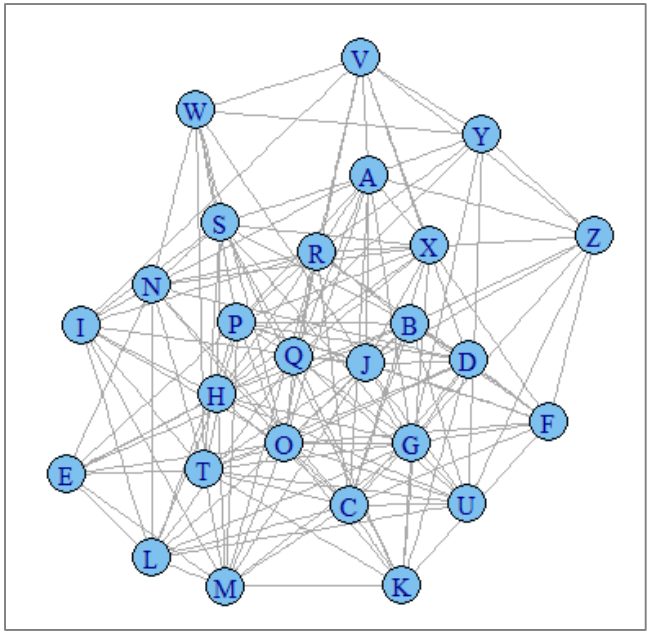


«Zeig mir deine Bücher und ich sag dir wer du bist.»

- Analyse von Büchern, Berichten, Webseiten etc.
- Beziehungsanalyse zwischen Textquellen (Assoziation)
- Beziehungsanalyse zwischen Textquellen und Eigentümer (Relation)

Social-Network-Analysen

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A	1	0	1	0	0	0	1	0	0	1	0	0	0	1	1	1	1	1	1		0	1	0	1	1	1
B	0	1	1	1	1	1	0	1	0	1	0	0	0	0	1	1	0	1	1		1	0	0	0	0	1
C	1	1	0	1	0	1	1	1	0	0	1	1	1	1	1	0	0	1	0	1	1	0	0	1	0	0
D	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	1	0	1	0	1	0	1	0	0	0	0
E	0	1	0	0	1	0	0	1	0	0	0	1	1	1	1	1	1	0	0	1	0	0	0	0	0	0
F	0	1	1	1	0	0	1	0	0	1	1	0	0	0	1	1	0	1	0		0	0	0	1	0	1
G	1	0	1	1	0	1	0	1	0	1	1	1	1	0	1	0	1	0	1	1	1	0	0	0	1	1
H	0	1	1	0	1	0	1	1	1	0	0	1	1	0	0	0	0	1	1		1	0	0	1	1	0
I	0	0	0	0	0	0	0	1	0	1	0	1	1	1	1	0	0	1	1	1	0	1	0	0	0	0
J	1	1	0	1	0	1	1	0	1	0	0	0	1	0	1	1	1	0	0	1	1	0	1	1	0	1
K	0	0	1	1	0	1	1	0	0	0	0	0	1	0	1	0	1	0	1	1	0	0	0	1	0	0
L	0	0	1	0	1	0	1	1	1	0	0	1	1	1	1	1	1	0	0		1	0	0	0	0	0
M	0	0	1	0	1	0	1	1	1	1	1	1	0	0	1	1	0	1	0	1	0	0	0	0	0	0
N	1	0	1	1	1	0	0	0	1	0	0	1	0	0	1	0	0	1	1	1	0	0	1	1	0	0
O	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	0	0	0
P	1	1	0	1	1	1	0	0	0	1	0	1	1	0	1	0	1	1	0	1	0	0	1	1	0	0
Q	1	0	0	0	1	0	1	0	0	1	1	1	0	0	1	1	0	1	1		1	1	0	1	0	0
R	1	1	1	1	0	1	0	1	1	0	0	0	1	1	1	1	1	0	0		0	1	0	1	1	0
S	1	1	0	0	0	0	1	1	1	0	1	0	0	1	0	0	1	0	0	1	0	0	1	1	0	0
T	0	0	1	1	1	0	1	0	1	1	1	0	1	1	1	1	0	0	1	1	1	0	1	0	0	0
U	0	1	1	0	0	0	1	1	0	1	0	1	0	0	1	0	1	0	0	1	0	0	0	0	1	1
V	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1	1	0		0	0	1	1	1	1
W	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	1	0	0	1	1	0	1	1	0	1	0
X	1	0	1	0	0	1	0	1	0	1	1	0	0	1	0	1	1	1	1		0	1	0	0	0	1
Y	1	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0		1	1	1	0	0	1
Z	1	1	0	0	0	1	1	0	0	1	0	0	0	0	0	0	0	0			1	1	0	1	1	1

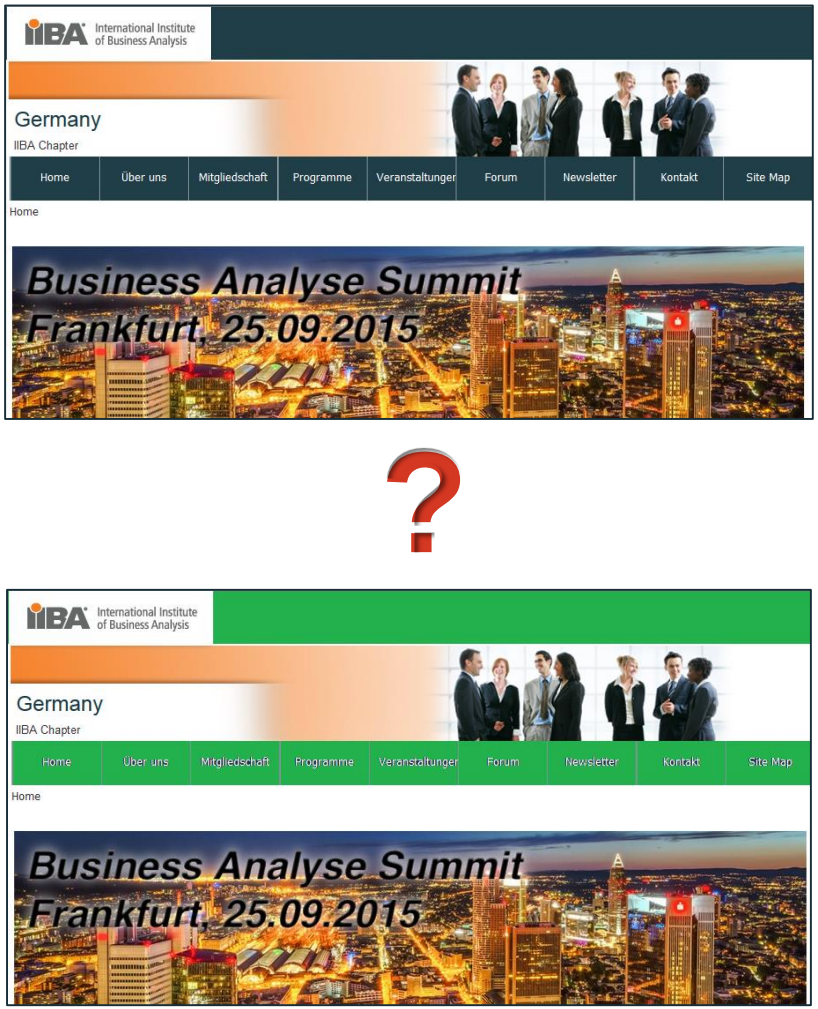


«Sag mir wer du bist und ich sag dir wer deine Freunde sind.»

- Kontaktanalysen (z.B. Email)
- Social-Media-Analysen (z.B. Twitter)
- Multidimensionale Cluster-Analysen

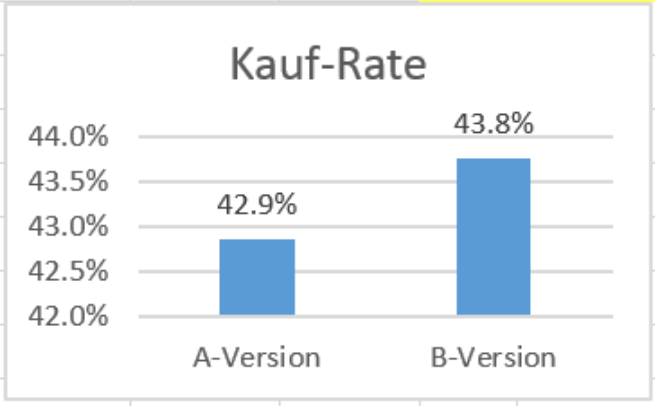
A/B Testing

- Beispiel: A/B-Test bei Internet-Auftritten
- Bewirken A-Version und B-Version unterschiedliche Anmeldezahlen?
- Tests sind aufwändig und teuer
- Repräsentative Stichproben sind aussagekräftiger
- Rückschluss von Stichprobe (Smart Data) auf Grundgesamtheit (Big Data) ist mit «Irrtums-Wahrscheinlichkeit» behaftet, bezüglich der Aussage, dass A-Version und B-Version sich statistisch signifikant unterscheiden

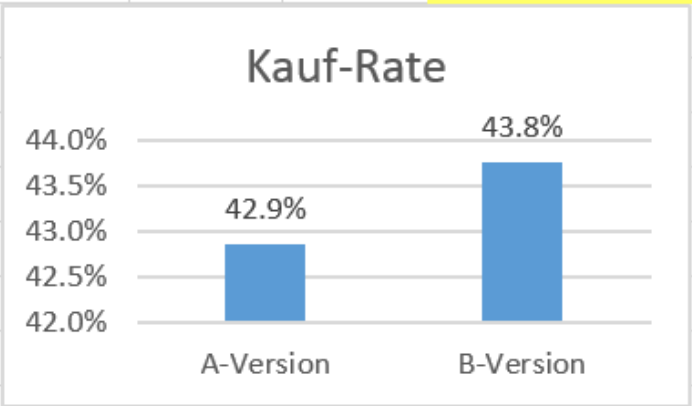


A/B Testing

	Kauf	Nein		% Kauf
A-Version	6	8	14	42.9%
B-Version	7	9	16	43.8%
	13	17	30	96.1%
			N	Irrtums-WS

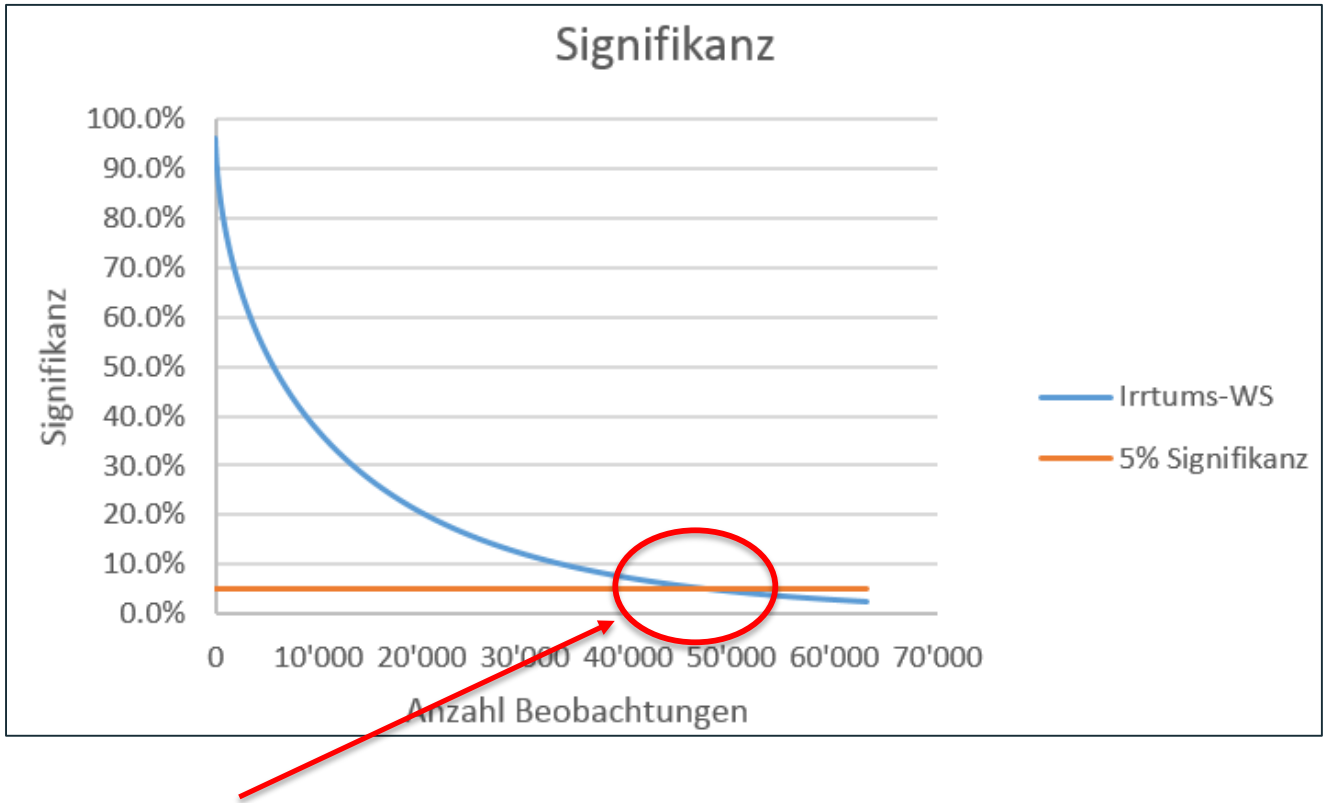


	Kauf	Nein		% Kauf
A-Version	6'000	8'000	14'000	42.9%
B-Version	7'000	9'000	16'000	43.8%
	13'000	17'000	30'000	11.9%
			N	Irrtums-WS



- Bei 30 Beobachtungen (links) ist die Irrtums-WS 96.1%, dass sich die zwei Anmelderaten (42.9% und 43.8%) signifikant unterscheiden.
- Bei 30'000 Beobachtungen (rechts) ist die Irrtums-WS nur noch 11.9%, dass sich die zwei Anmelderaten (42.9% und 43.8%) signifikant unterscheiden.

A/B Testing



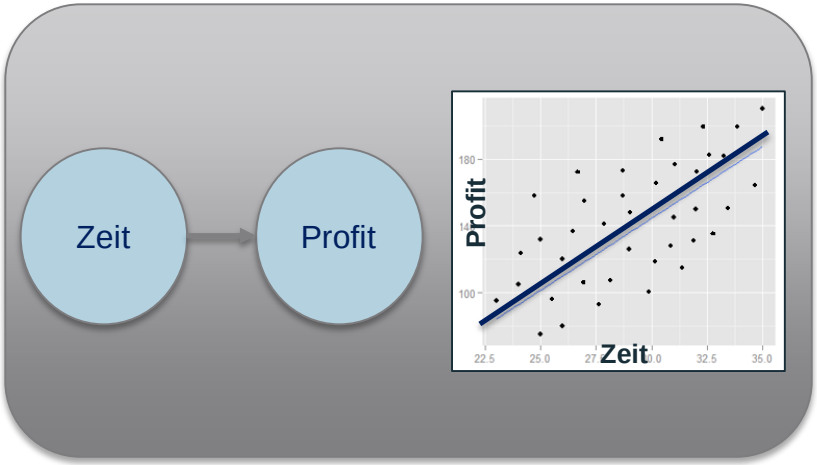
- Bei 47’340 Beobachtungen wird der Unterschied zwischen den zwei Anmelderaten (42.9% und 43.8%) statistisch signifikant.
- Mit der Masse des «Big Data» wird (leider) alles statistisch signifikant!
- Fazit: Lieber ausgewählte Smart Data als erdrückende Big Data.

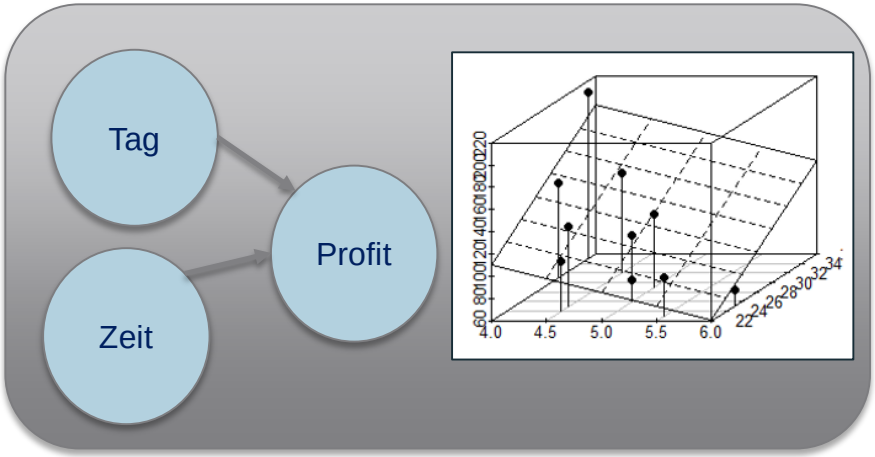
Knowledge-Mapping

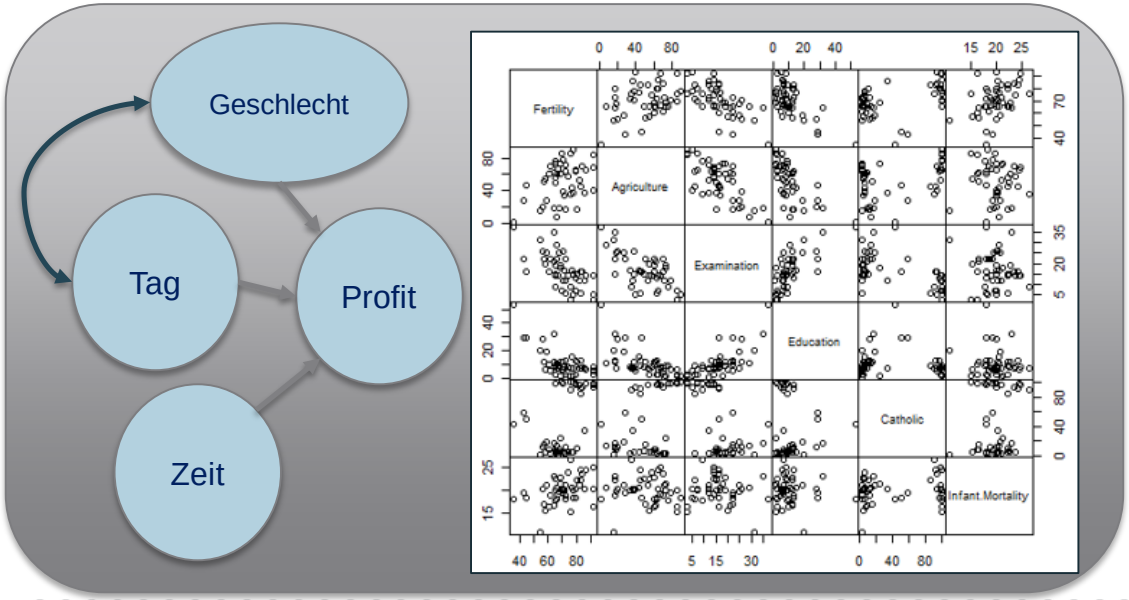
Eingabe in Google: [«Je mehr desto» Frankfurter Allgemeine]

- Je mehr Volumen [an der Börse], desto kleiner die kurzfristigen Ausschläge
- Personalgespräche Je höher der Posten, desto mehr Schauspiel
- Digitaldebatte: Je größer die Mythen vom Netz, desto kleiner der Mensch
- Je mehr Feinstaub durch die Luft schwirrt, desto schlechter spielen die Profis
- Je mehr Muskel, desto besser: Das Streben nach diesem Körperbild kann krankhafte Züge annehmen.
- Je mehr man darüber liest, desto weniger weiß man, woran man überhaupt noch glauben kann und soll

Knowledge-Mapping

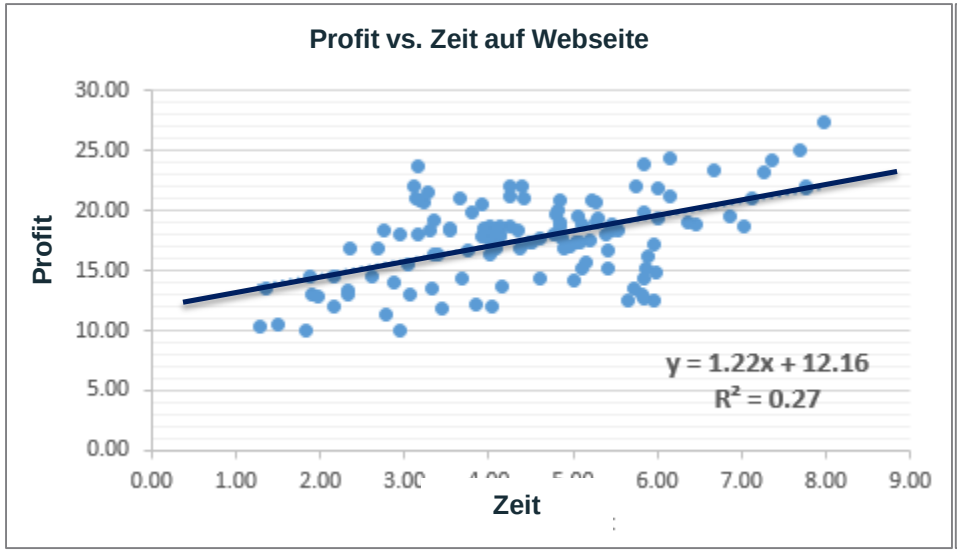


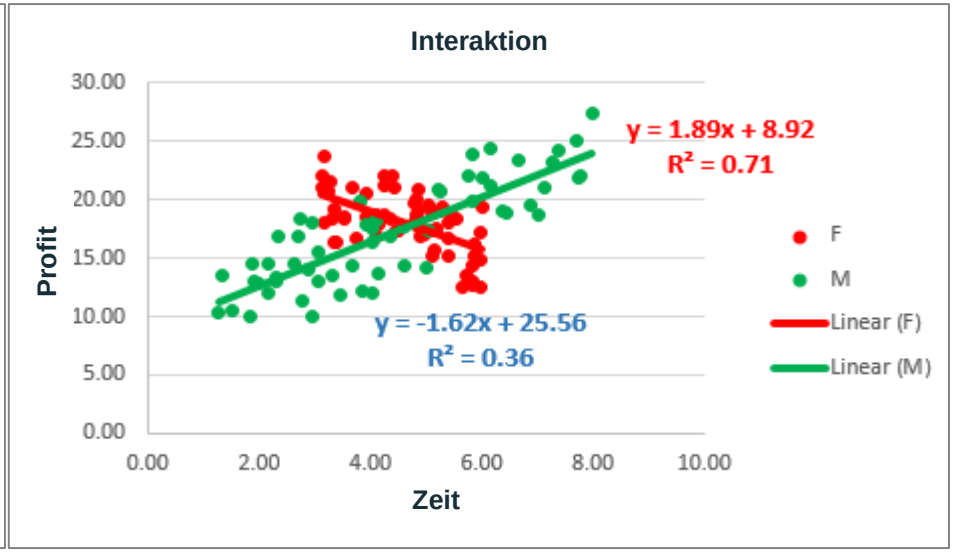


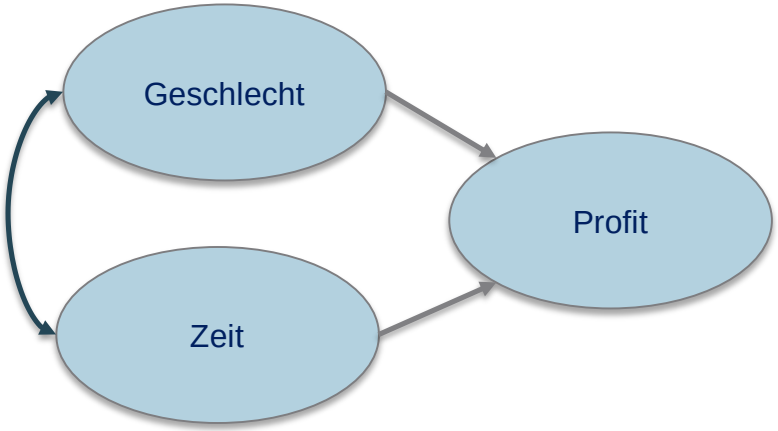


Achtung auf:
- Korrelation vs. Kausalität
- Multi-Ko-Linearität
- Partielle Korrelation
- Interaktions-Effekt

Knowledge-Mapping



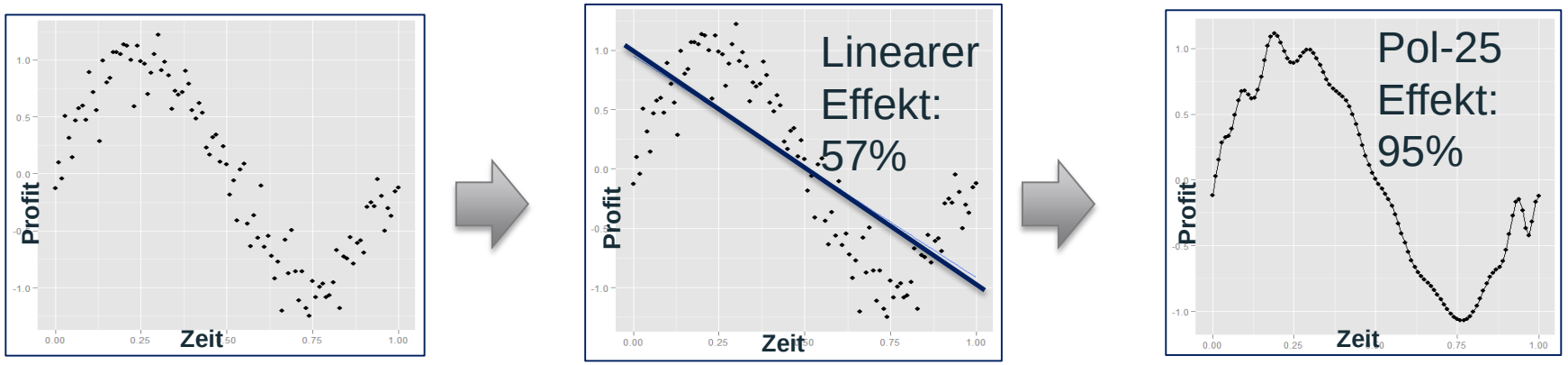




Interaktion:

- Positive Korrelation zwischen Zeit und Profit
- Partielle Korrelation ist unklar:
  - Positive Korrelation für Männer
  - Negative Korrelation für Frauen

Machine-Learning

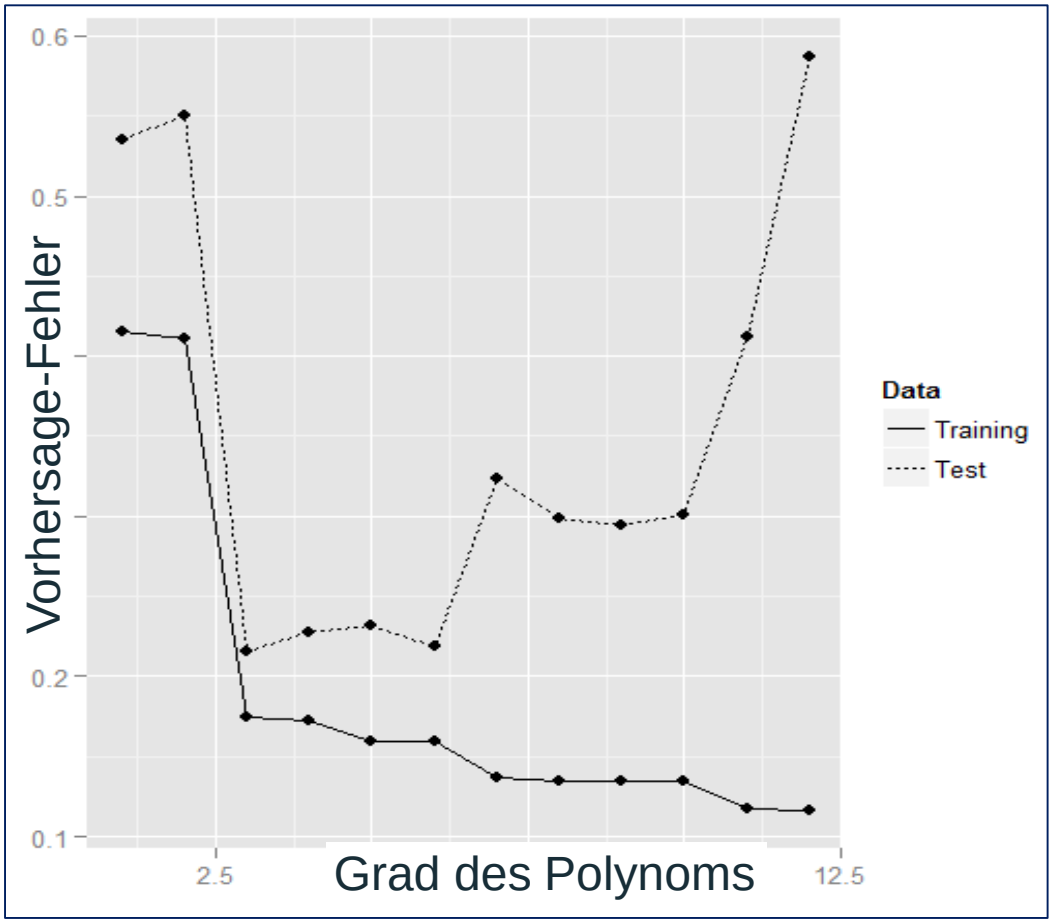


Machine-Learning: Der Computer optimiert das Modell

- Verweildauer (Zeit) auf einer Internet-Seite und Profit korrelieren auf einer Sinus-Kurve: zuerst steigt der Profit an, mit steigender Verweil-Zeit, dann sinkt er wieder, um dann erneut anzusteigen.
- Lineare Effekt-Stärke ist nur 57% ($K = a_1*Z + a_0$)
- Polynom-25 Effekt-Stärke ist 95% ($K = a_1*Z^{25} + a_2*Z^{24} + \ldots + a_{25}*Z + a_0$)
- Was für ein Polynom sollen wir nehmen, um möglichst gute Vorhersagen des Profits machen zu können?
- Predictive-Modelling teilt Daten in zwei Sets ein: Eines für die Modell-Erstellung (Training) und das andere für die Validierung (Test).

Machine-Learning

- Das mit dem Training-Set erstellte Modell wird auf die Test-Daten angewendet.
- Beide Vorhersage-Fehler sinken bis Grad 3 des Polynoms.
- Der Vorhersage-Fehler für das Test-Set steigt dann stark an.
- Die «Schere» zwischen Modell und Wirklichkeit öffnet sich.
- Wir haben ein «Over-Fitting» und Modellieren nur noch «Kaffeesatz».
- Das Modell ist nicht mehr relevant, da die Zitrone schnell mal ausgedrückt ist!
- Fazit: Daten nicht in den Schraubstock spannen!



Impact auf die Business Analysis

1. Daten-Filter statt Daten-Galaxien
2. Neue Analyse Methoden
3. Reaktive-dynamische Analysen

1. Daten-Filter statt Daten-Galaxien

- Um die 4V der Big Data zu optimieren, müssen Daten-Filter definiert werden (für die Sammlung und für die Löschung von Daten).
- Big Data Strategie definieren (für eine Data-Driver Organisation) (Scheinwerfer-Prinzip statt Garbage-Can-Prinzip) und mit dem Business-Modell (z.B. CRM Kundeninteraktion) verlinken.
- Big Data Infrastruktur soll optimiert werden (z.B. Cloud-Lösung, mobile und agile Lösungen).
- Anstelle von hieroglyphischen log-Files, sollen gezielte Daten-Sammlungen und Daten-Modelle definiert werden.
- Dynamik der Daten und subjektive, interaktive Aspekte der Daten sollen miteinbezogen werden.
- Daten-Schutz und Daten-Ethik muss geklärt werden.

2. Neue Analyse Methoden

- Predictive Analytics:
  - Zur Segmentierung und Profilierung von Zielgruppen
  - Basierend auf Validierung der Modelle, anstatt nur auf Signifikanz und Qualität.

- Web und Mobile Analytics:
  - Geo-Location Analytics (Stau-Management)
  - Customer-Behavior Analytics (Datenschutz!)

- Social Network and Behavior Analytics:
  - Struktur-Modelle (Faktorenanalysen, Bäume (CART) etc.)
  - Geo-Location und Visual Analytics

- Machine Learning und Collective Intelligence:
  - Bootstrapping, Algorithmen, Simulation

3. Reaktive-dynamische Analysen

- Mit dem grossen Datenfluss kann man «Echtzeit-Modelle» erstellen:
  - Diese passen sich automatisch mit neuen Daten an
  - Sie können auch subjektive Elemente einbeziehen

- Beispiel: High-Freqeuncy Trading, Auktionen-basierte Offerten (Hotel, Flugpreise) mit variierenden (subjektiven) Mindest-Offerten und variablen (vom Markt bestimmten) Verkaufs-Preisen

- Offene Fragen:
  - Wie beeinflusst der (variable) Mindestpreis die schlussendlichen Verkaufs-Preise?
  - Was sind die Einflussfaktoren für die Auktions-Preise (Saison, Destination etc.)?
  - Welche Kunden-Typen nehmen an Auktionen statt, wer verschliesst sich solchem Kaufverhalten?

Fazit

- Besser «smart» als «big»
- Integration von objektiven Daten und subjektiven Entscheidungspräferenzen
- Daten-driven Business Modell für Prozesse und Infrastruktur
- (Unternehmens-)übergreifende Zusammenarbeit in Virtual Teams
- Analytisches Know-How nicht in die Cloud outsourcen