Institut für Finanz- und Aktuarwissenschaften

Data Analytics

Was sind diese Bäume und Wälder?
Und viel wichtiger: was kann man mit denen machen?

- Lukas Hahn

- DAV vor Ort, Stuttgart
- 25. September 2018

Data Analytics
Agenda

Data Analytics: Was ist das eigentlich?

Ein Exkurs zu Bäumen und Wäldern

Was fangen wir damit an?

Institut für Finanz- und Aktuarwissenschaften

Data Analytics: Was ist das eigentlich?

Eine Definition

Data Analytics

… is the extensive use of data, statistical and quantitative analysis, explanatory and predictive models, and fact-based management to **drive decisions and actions**.

Davenport, Thomas and, Harris, Jeanne (2007). Competing on Analytics. O'Reilly.

… is the discovery, interpretation, and communication of **meaningful patterns** in data.

Englischer Wikipedia-Eintrag zu „Analytics", Stand 17.04.2018

- Versicherungsunternehmen besitzen **große Datenmengen**, die zahlreiche Informationen z.B. zu Kunden und Schäden enthalten. Data Analytics beinhaltet die intelligente Informationsgewinnung aus solchen Daten und die praktische Umsetzung der daraus gewonnenen Erkenntnisse.
- Wesentliche Prozessschritte sind die **Konkretisierung** der Zielsetzung und Datenanforderung, die **technische Datenanalyse**, die kontextbasierte **Auswertung,** die **Interpretation und Kommunikation** von gewonnenen Erkenntnissen sowie die daraus abgeleitete **Entscheidungsfindung** und **Umsetzung**.
- Neben klassischen Ansätzen der Datenanalyse kommen dabei vermehrt **Methoden des Machine Learnings** zur Anwendung (**Advanced Analytics**).

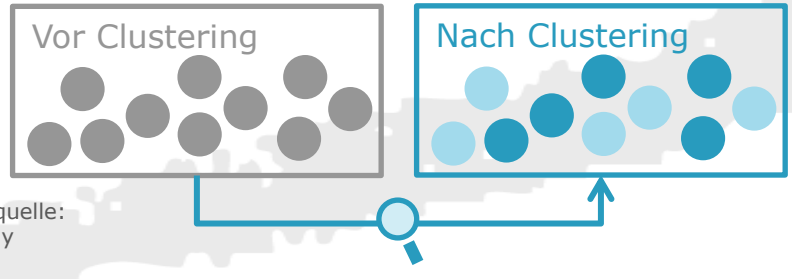
Bildquelle: Pixabay

Data Analytics: Was ist das eigentlich?

Typische Fragestellungen

Data Mining

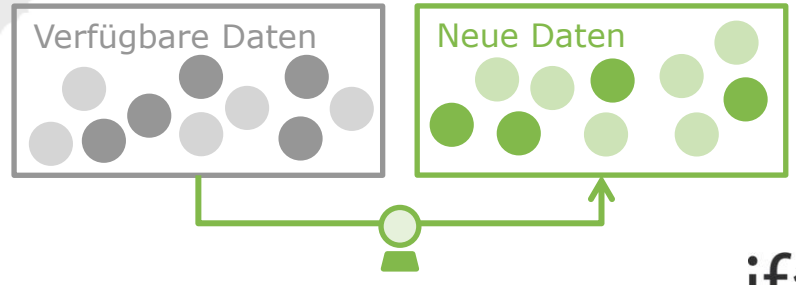
- Identifikation komplexer **Muster**
  - Ziel: Wissensgenerierung, z.B. Clustering
  - Aufgabe: Identifizierung systematischer Zusammenhänge in vorhandenen Daten
  - Beispiel: Kunden oder Vertriebspartner gruppieren, Auffälligkeiten entdecken



Bilderquelle: Pixabay

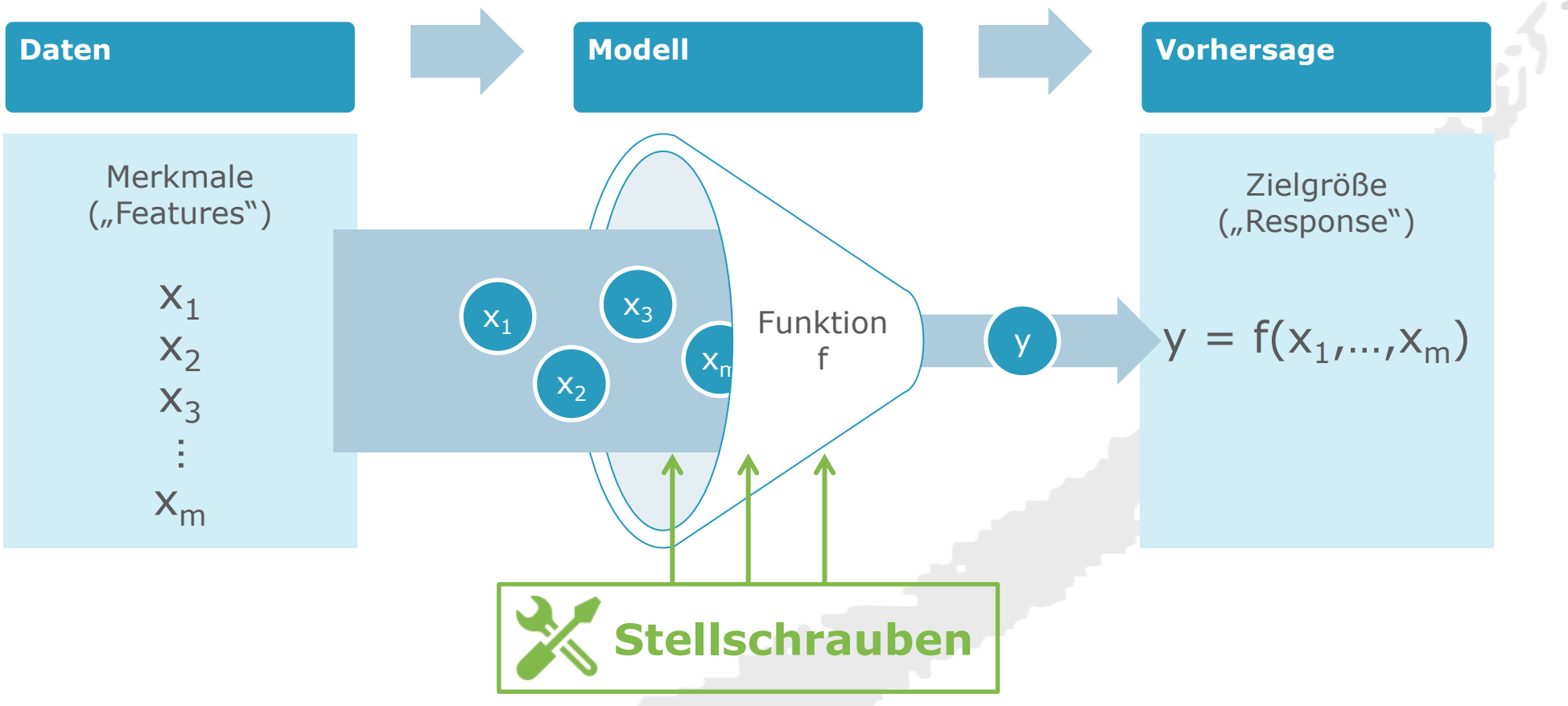
Predictive Modelling

- Bestmögliche individuelle **Vorhersage**
  - Ziel: optimale Entscheidungsfindung
  - Aufgabe: Identifizierung systematischer Vorhersageregeln für neue Daten
  - Beispiel: Storno vorhersagen, Schäden projizieren



Data Analytics: Was ist das eigentlich?

Was genau ist ein Modell?



- Ein Data-Analytics-Modell ist eine mathematisch-statistisch geschätzte **Funktion**, die den eingehenden Daten (Merkmalen) eine Vorhersage (Zielgröße) zuordnet.
- Über **Stellschrauben** der Funktion wird in Abhängigkeit der Datenbeschaffenheit und der unternehmerischen Zielsetzung die bestmögliche Vorhersage modelliert.

Data Analytics: Was ist das eigentlich?

In der Theorie

- Mathematisch schätzen wir also eine (abstrakte) Funktion f mit $Y = f(X_1, X_2, \ldots, X_m) + \varepsilon$, die die realen Zusammenhänge („Muster") zwischen $X_1, X_2, \ldots, X_m$ und Y beschreibt.
  - f macht die „bestmögliche" Aussage von $X_1, X_2, \ldots, X_m$ über Y und ist im Mittel korrekt.
  - ε beschreibt die verbleibenden Abweichungen, die nicht mit $X_1, X_2, \ldots, X_m$ erklärbar sind.
- Unsere Schätzung $\hat{f}$ beruht auf gewissen Annahmen an f um sie gut bestimmen zu können.
  - $\hat{f}$ soll f möglichst gut approximieren, indem wir eine Verlustfunktion $L(Y, \hat{f}(X))$ minimieren.
- Theoretisch ist das nicht neu, denn die statistische Schätzung von Funktionen und auch „moderne" Verfahren existieren schon lange (neuronale Netze: 1950er, baumbasierte Verfahren: 1980er, …).
- Neu sind aber vor allem
  - die Menge verfügbarer bzw. sinnvoll erfasster, gespeicherter und zusammengeführter Daten,
  - die Rechenperformance um diese Daten mit verschiedenen komplexen Methoden zügig oder in Echtzeit auszuwerten und die Ergebnisse zu analysieren.

Wir sind somit insbesondere in der Lage mehr **Stellschrauben** zu nutzen:

- Muster in unseren Daten durch komplexere Verfahren zu identifizieren und
- die Mustererkennung durch sog. „Tuning" der Modelle zu optimieren.

Data Analytics
Agenda

Data Analytics: Was ist das eigentlich?

Von Bäumen und Wäldern

Was fangen wir damit an?

Institut für Finanz- und Aktuarwissenschaften

Von Bäumen und Wäldern

Einführung: Entscheidungsbäume

Bildquelle:
http://www.freeiconspng.com
31.03.2017

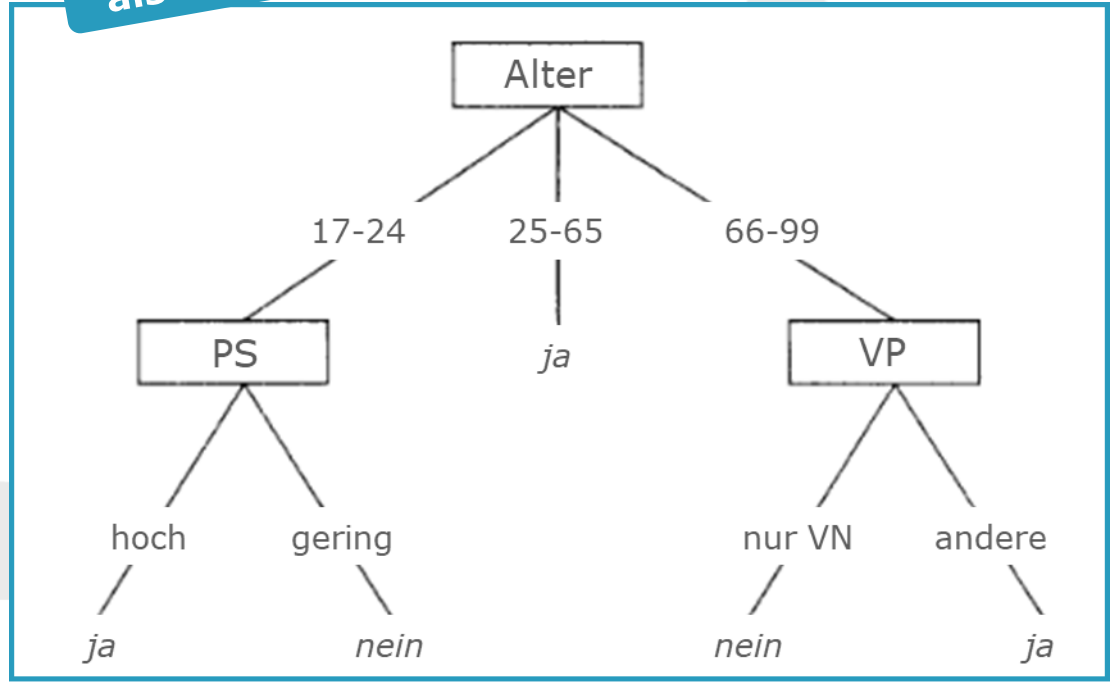
Beispiel zur Veranschaulichung

- Fragestellung: Bei welchen Verträgen gab es einen Schaden in der Kfz-Haftpflichtversicherung?
- Entscheidungsregeln: Tarifmerkmale

als Daten

Alter	PS	VP	Schaden?
25-65	hoch	nur VN	ja
25-65	gering	nur VN	ja
25-65	gering	andere	ja
17-24	hoch	nur VN	ja
17-24	hoch	andere	ja
17-24	gering	nur VN	nein
17-24	gering	andere	nein
66-99	hoch	nur VN	nein
66-99	gering	nur VN	nein
66-99	hoch	andere	ja
66-99	gering	andere	ja

als Baum



Von Bäumen und Wäldern

Klassifikations- und Regressionsbäume

Induktion eines Baums

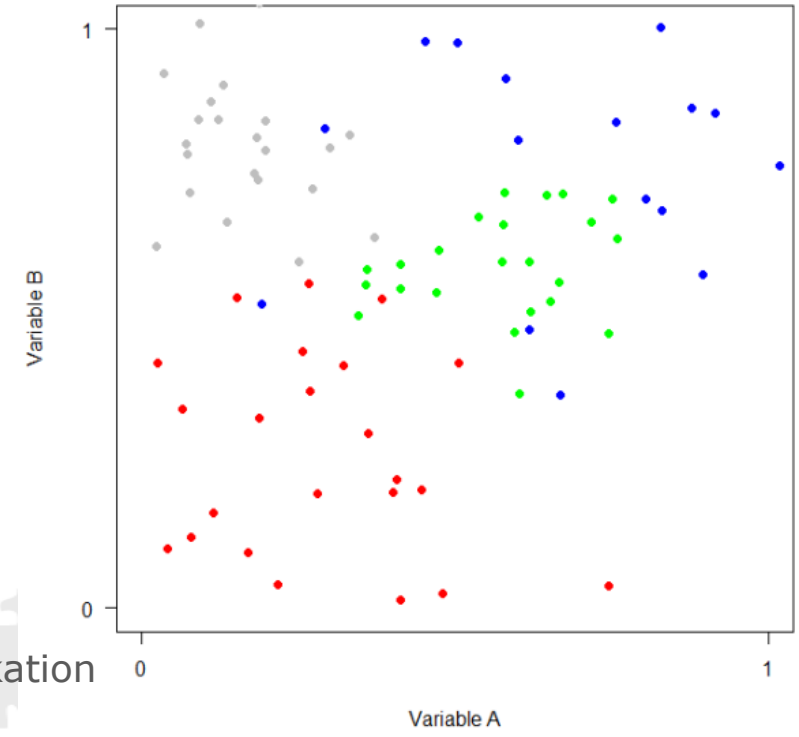
- Ausgangssituation: Trainingsdaten mit bekannter Klassifizierung / Zielgröße
- **Baum wachsen lassen (Growing)**
  - Rekursives Top-Down-Prinzip
    - Iterative Vorgehensweise
      - Welches Attribut unterscheidet am besten?
      - Welche Entscheidungsregel unterscheidet am besten?
      - Auswahl via mathematischer Bestimmungsmaße
    - Top-Down
      - vom Allgemeinen (Wurzel) zum Konkreten (Blätter)
      - basierend auf bisherigem Teilbaum (von oben nach unten lesen)
    - Rekursiv
      - Wiederanwendung derselben Logik bei jeder Verästelung
- **Baum stutzen (Pruning)**
  - Early stopping (Pre-Pruning): Stoppregel beim Wachsen
  - (Post-)Pruning: Zurückschneiden eines vollständig gewachsenen Baums

Von Bäumen und Wäldern

Beispiel: Klassifikationsbaum zur Kundensegmentierung

illustrativ

- **Aufgabe:** Kunden anhand zweier Merkmale klassifizieren für geeignetes Zielgruppenmarketing
  - Zwei (normierte) Merkmale:
    - Anbindungsdauer des Kunden (Variable A)
    - Summe jährlicher Beiträge (Variable B)
  - Zielgröße mit vier Kategorien: Veränderung der Kundenbeziehung im nächsten Jahr
    - **rot**: Verschlechterung (z.B. durch Storno)
    - **grün**: Verbesserung innerhalb einer Sparte (z.B. neue Verträge oder Aufstockung)
    - **blau**: spartenübergreifende Verbesserung (z.B. Vertragsabschluss in weiterer Sparte)
    - **grau**: keine Veränderung
- **Performancekriterium:** möglichst geringe Fehlklassifikation
- **Algorithmus („Lerner"):** Klassifikationsbaum

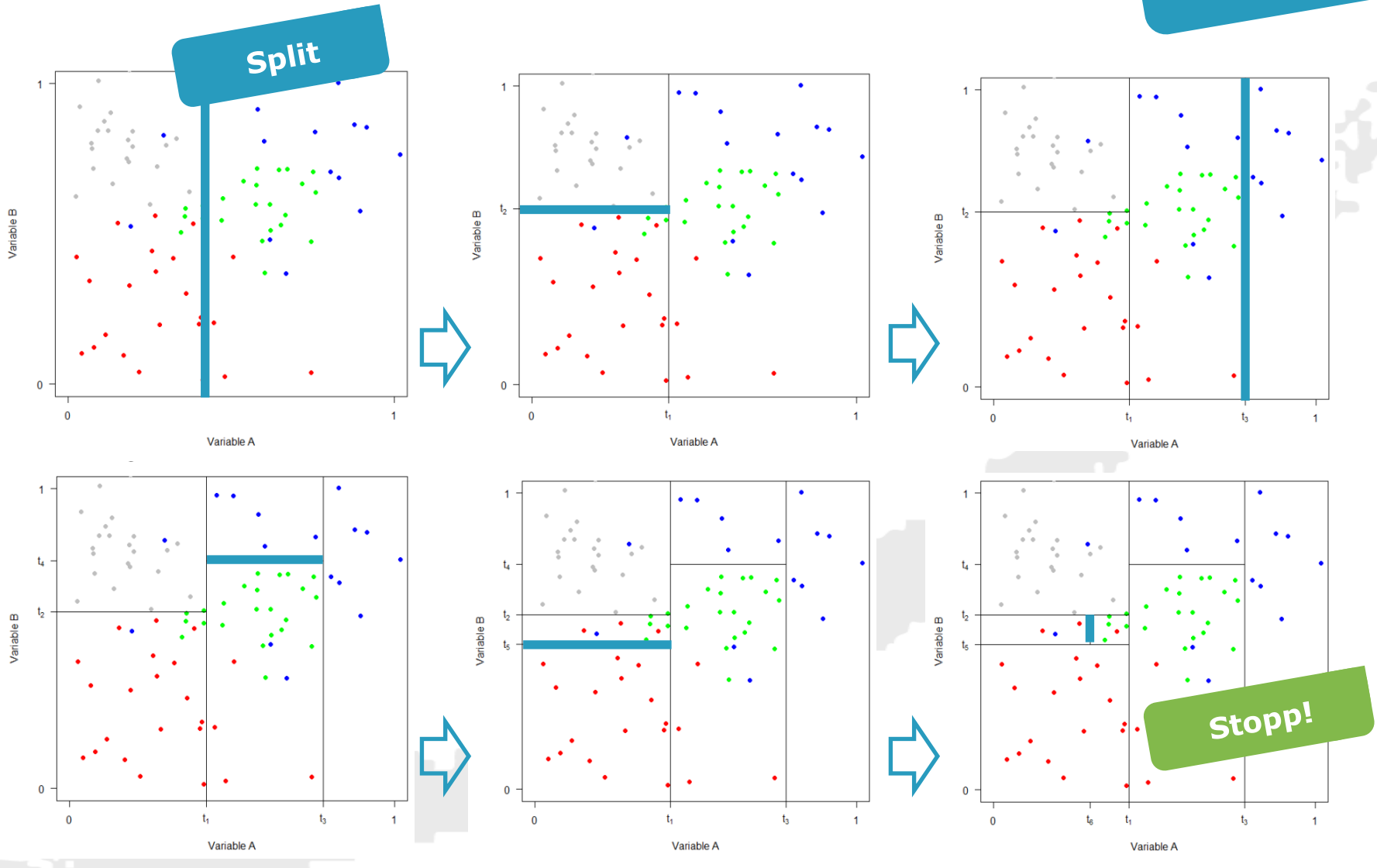


Bildquelle: Weinhold (2014), Analyse und Anwendung von Entscheidungsbäumen zur Fehlererkennung im Gebäudebetrieb.

Von Bäumen und Wäldern

Beispiel: Klassifikationsbaum zur Kundensegmentierung

illustrativ

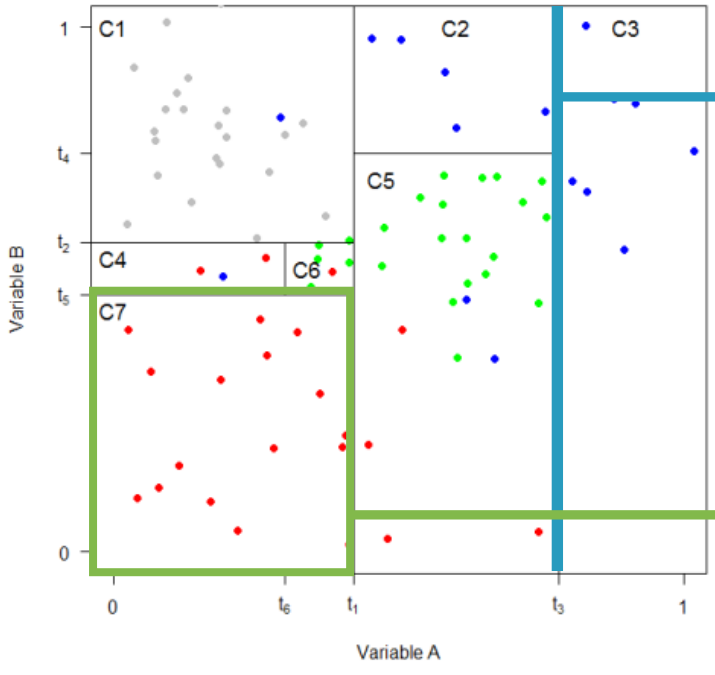


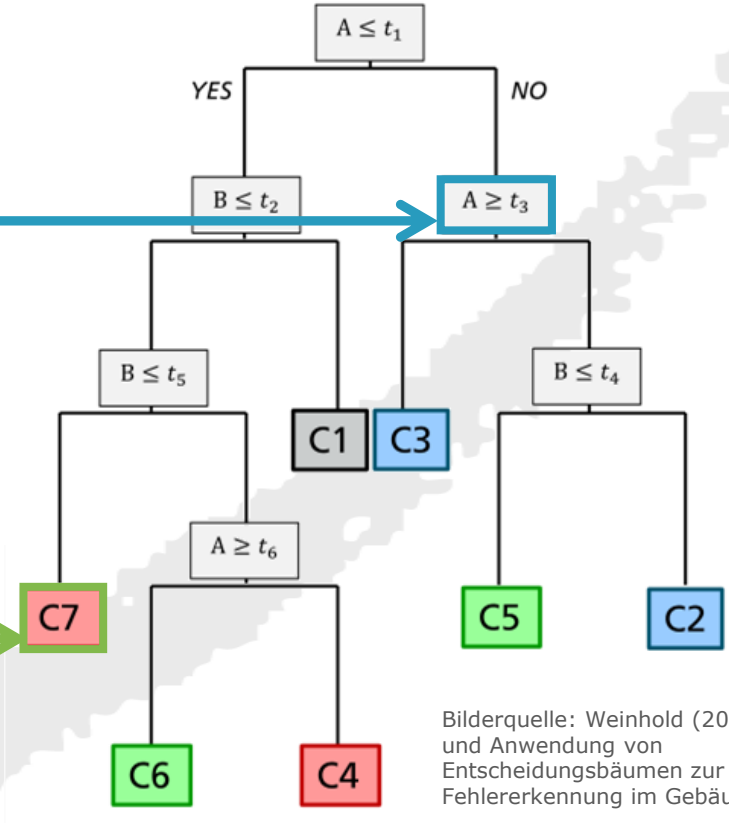
Bilderquelle: Weinhold (2014), Analyse und Anwendung von Entscheidungsbäumen zur Fehlererkennung im Gebäudebetrieb.

Von Bäumen und Wäldern

Beispiel: Klassifikationsbaum zur Kundensegmentierung

illustrativ





Bilderquelle: Weinhold (2014), Analyse und Anwendung von Entscheidungsbäumen zur Fehlererkennung im Gebäudebetrieb.

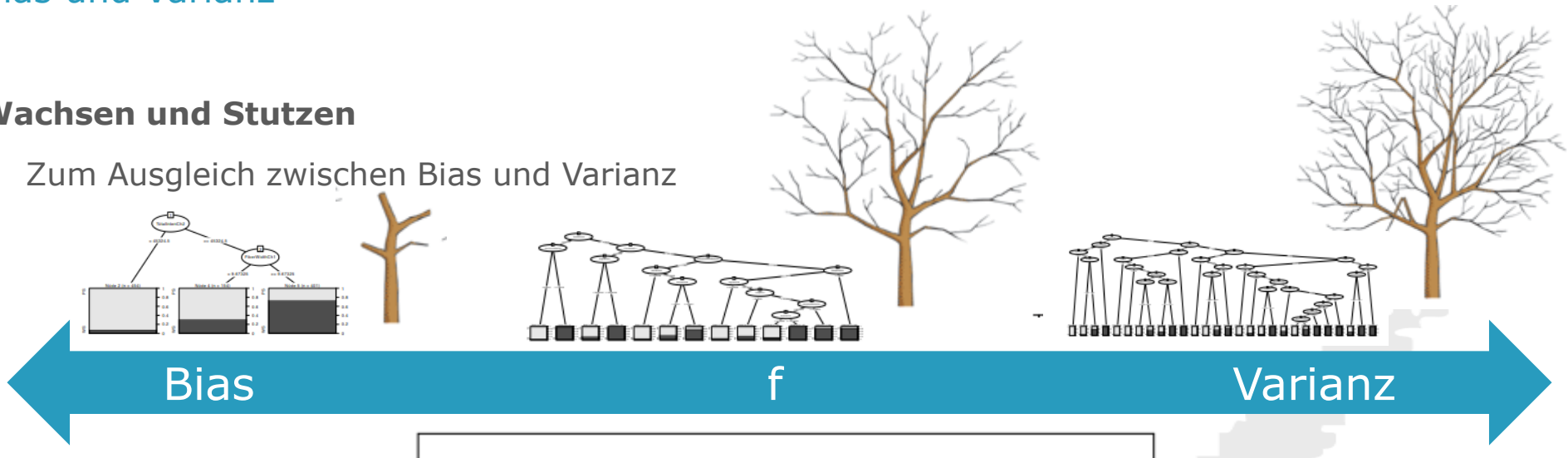
Von Bäumen und Wäldern

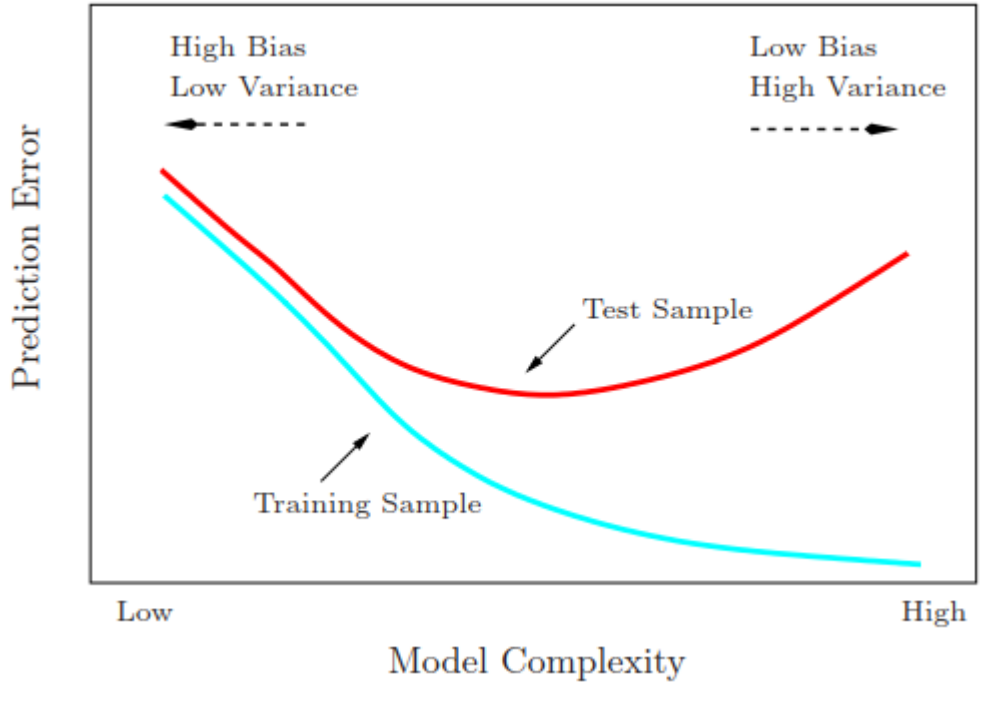
Bias und Varianz

Bilderquelle:
https://www.thetreecenter.com/how-and-when-to-prune-fruit-trees, 31.03.2017

Wachsen und Stutzen

- Zum Ausgleich zwischen Bias und Varianz





Bildquelle:
Hastie et al. (2009). The Elements of Statistical Learning – Data Mining, Inference, and Prediction

Von Bäumen und Wäldern

Bias und Varianz

Bias und Varianz

- Wir suchen ein Modell f mit minimalem Fehler, z. B. mit minimaler quadratischer Abweichung.
  - In perfekter Modellwelt (unendliche Daten) ist dies nur der **unvermeidbare Fehler**.
  - Für endliche Stichproben („Sample") verbleiben die Fehlerkomponenten **Bias und Varianz**.

$$E_{Sample|X}\left(\left(Y-\hat{f}(X)\right)^2\right) = \left(E_{Sample|X}\left(\hat{f}(X)\right)-f(X)\right)^2 + Var_{Modell}(\varepsilon) + E_{Sample|X}\left(\left(\hat{f}(X)-E_{Sample|X}\left(\hat{f}(X)\right)\right)^2\right)$$

Abweichung der mittleren Modellvorhersagen von der Realität	Unvermeidbarer Fehler	Variabilität in der Modellvorhersage durch die Stichprobe

- Herausforderung: simultane Minimierung von (Bias² + Varianz)
  - Erhöhung der Stichprobengröße (ist aber i.d.R. fix)
  - Modellwahl (ist aber i.d.R. eingeschränkt)

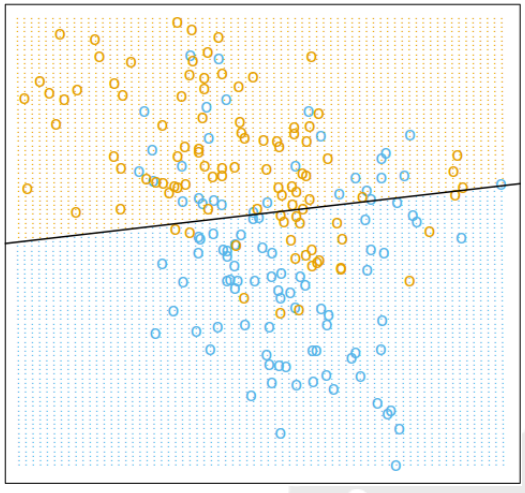
Es bleibt ein Tradeoff zwischen Bias und Varianz.

Von Bäumen und Wäldern

Bias und Varianz

Bias

- Hoher Bias, geringe Varianz: einfache Modelle mit globalen Annahmen, z. B. Linearität
  - Modelle neigen zur Unteranpassung („**underfitting**").
  - Systematische Muster in der Grundgesamtheit bleiben unerkannt.



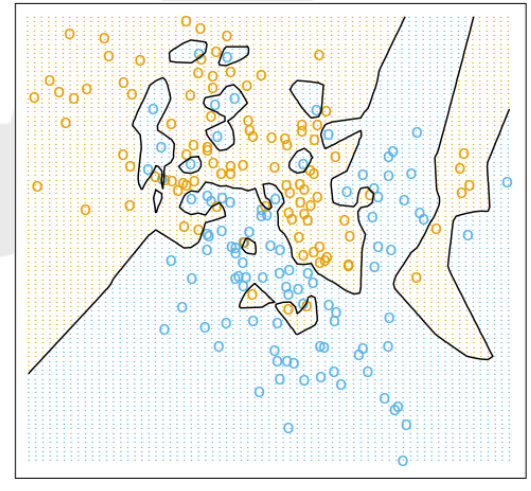
Bildquelle:
Hastie et al. (2009). The Elements of Statistical Learning – Data Mining, Inference, and Prediction

Von Bäumen und Wäldern

Bias und Varianz

Varianz

- Geringer Bias, hohe Varianz: komplexe Modelle mit lokalen Annahmen, z. B. nächste Nachbarn
  - Modelle neigen zur Überanpassung („**overfitting**").
  - Anpassung des Modells an das Rauschen der Stichprobe und nicht an die Struktur in den Daten



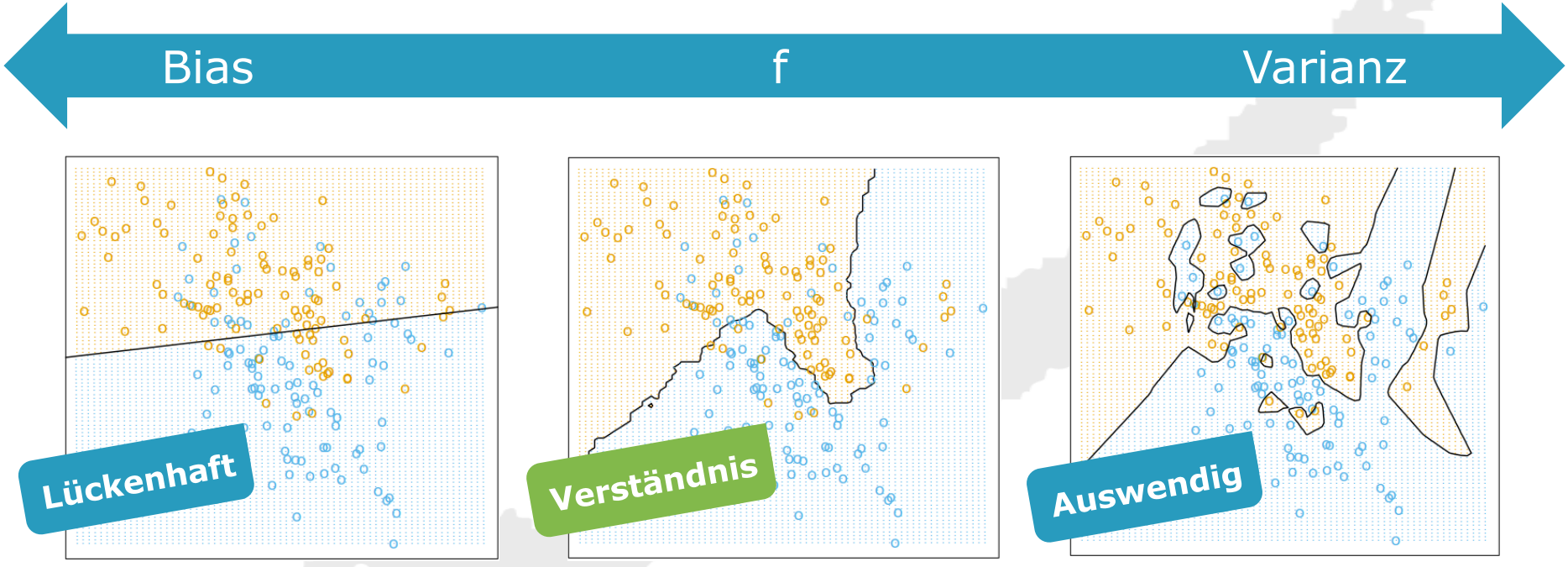
Bildquelle:
Hastie et al. (2009). The Elements of Statistical Learning – Data Mining, Inference, and Prediction

Von Bäumen und Wäldern

Bias und Varianz

Bias-Varianz-Dilemma

- Ziel ist ein **Kompromiss zwischen Bias und Varianz**:
  - Das Modell soll die systematischen Muster der Grundgesamtheit erfassen.
  - Das Modell soll das unsystematische Rauschen der Stichprobe unberücksichtigt lassen.



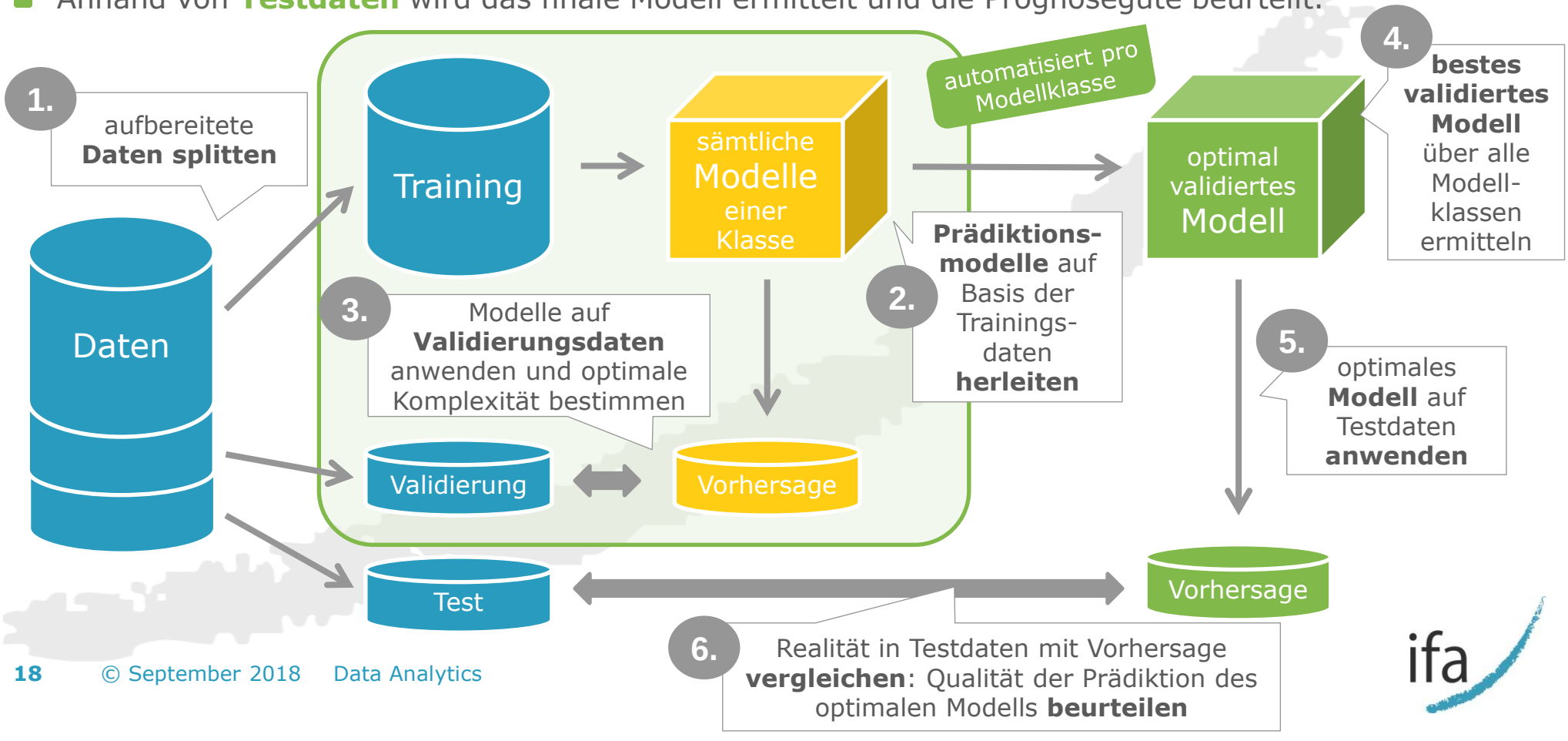
Bildquelle:
Hastie et al. (2009). The Elements of Statistical Learning – Data Mining, Inference, and Prediction

Von Bäumen und Wäldern

Training, Validierung und Test

Die Optimierung des Lernprozess erfolgt mit **Aufteilung der Daten** für Training, Validierung und Test:

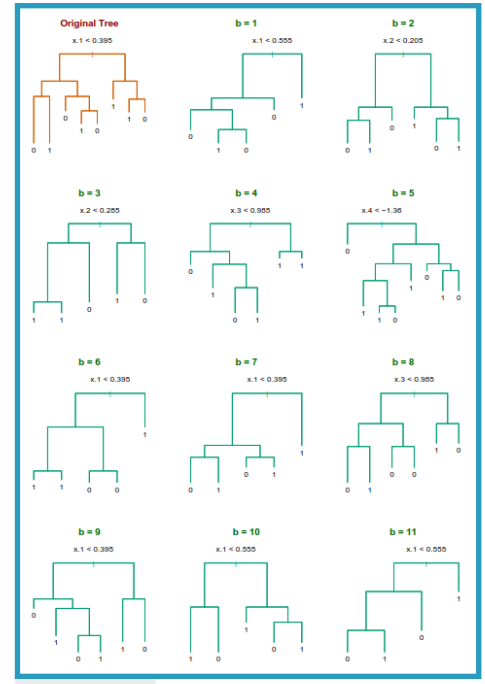
- Auf den Erfahrungen in den **Trainingsdaten** lernt jedes Modell (verschiedene Komplexitäten).
- Mittels Erfahrungen in den **Validierungsdaten** wird die optimale Komplexität pro Modell ermittelt.
- Anhand von **Testdaten** wird das finale Modell ermittelt und die Prognosegüte beurteilt.



Von Bäumen und Wäldern

Methoden des Ensemble-Learning

Bildquelle:
Hastie et al. (2009). The Elements of Statistical Learning – Data Mining, Inference, and Prediction



Methoden des Ensemble-Learning

- Grundidee:
  - einen bekannten Lernalgorithmus (z.B. CART) **mehrfach anwenden**
  - **„durchschnittliche" Vorhersage** als finales Modell verwenden
- Motivation:
  - Beobachtung: einzelne Modellinstanz tendiert zu Overfitting
  - Ansatz: durch Mittelung mehrerer Modellinstanzen die **Varianz senken** (bei konstantem Bias)
  - Ziel: bessere Vorhersagegüte des Ensemble der Modellinstanzen im Vergleich zur Einzelinstanz
- **Vorteil:** einen vergleichsweise „schwachen" Lernalgorithmus mit einfachen Mitteln stärken
  - keine alternativen Algorithmen oder Modelle notwendig
  - i.A. kein neuerliches Overfitting durch zu hohe Komplexität (Anzahl an Einzelinstanzen)
- **Nachteil:** Verlust der Interpretierbarkeit, höhere Rechenlaufzeiten

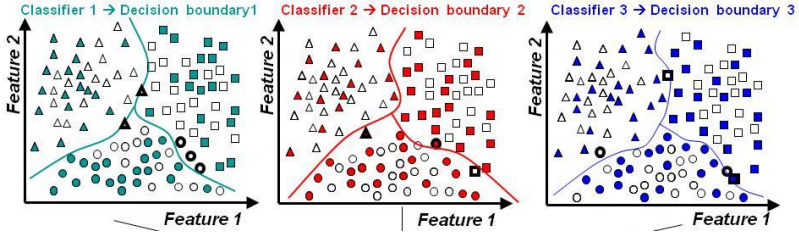
Von Bäumen und Wäldern

Random Forest

Bilderquellen:
Mercy (2012), Ensemble Learning and Model Selection, http://www.vias.org/tmdatanaleng/cc_linvsnonlin.html, 31.03.2017
Biodiversity and Climate Change Virtual Laboratory (2016), Random Forest, https://support.bccvl.org.au/support/solutions/articles/6000083217-random-forest#header-page3, 31.03.2017.

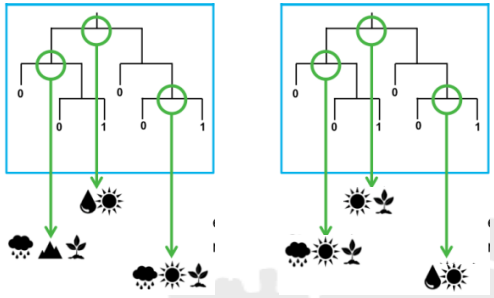
Bagging (Bootstrap Aggregating)

- Ziehe zufällig aus Beobachtungen.

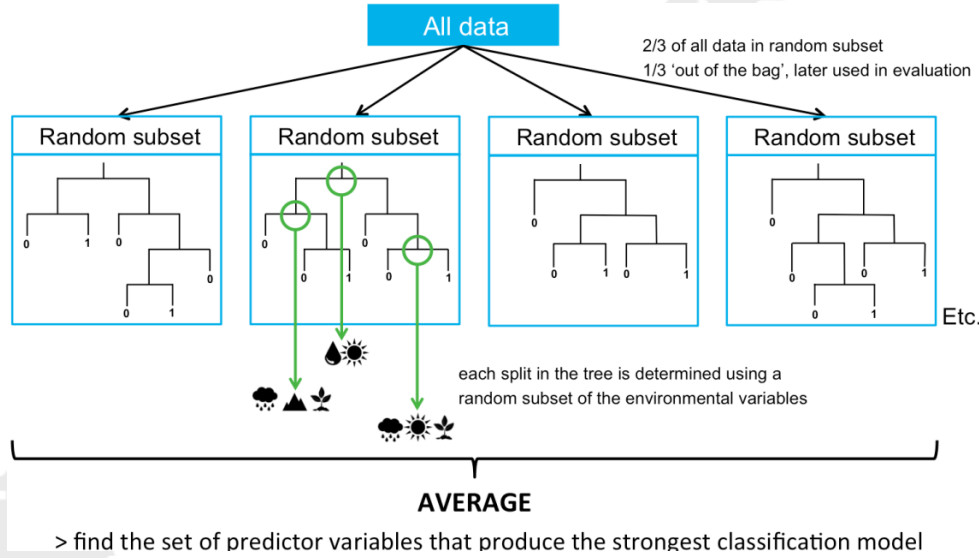


Random Subspace Method

- Ziehe zufällig aus Merkmalen pro Split.



Random Forest



> find the set of predictor variables that produce the strongest classification model

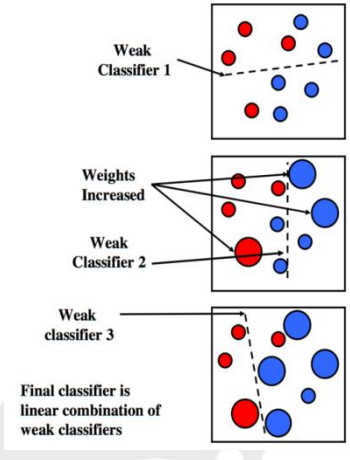
Von Bäumen und Wäldern
Boosting

Bilderquellen:
Jain (2016), Computer Vision – Face Detection, Vinsol, http://vinsol.com/blog/2016/06/28/computer-vision-face-detection, 31.03.2017.
Prettenhofer and Louppe (2014), Gradient Boosted Regression Trees
James et al. (2013), An introduction to statistical learning – with applications in R.

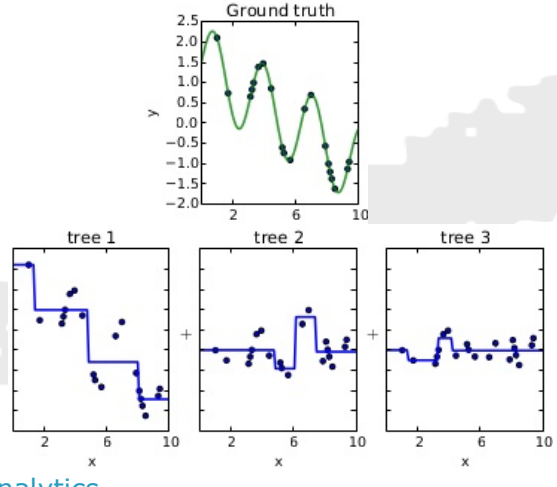
- Grundidee
  - wie bisher ein „Komitee" aus einzelnen Instanzen eines „schwachen Lerners" herleiten
  - Aber statt parallel werden die Instanzen beim Boosting sequentiell trainiert.
- Motivation
  - Fokus auf Verringerung des Bias: neue Instanz soll gezielt auf Beobachtungen trainiert werden, die von bisherigen Instanzen fehlerhaft vorhergesagt werden
  - alternativ: neue Instanzen anhand der Residuen des bisherigen Komitees trainieren

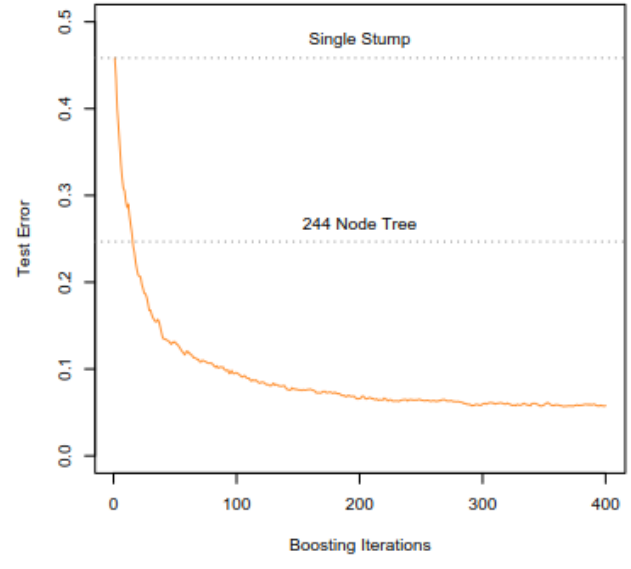
➔ **verschiedene Boosting-Algorithmen**

AdaBoost.M1



Gradient Boosting





Data Analytics

Agenda

Data Analytics: Was ist das eigentlich?

Ein Exkurs zu Bäumen und Wäldern

Was fangen wir damit an?

Institut für Finanz- und Aktuarwissenschaften

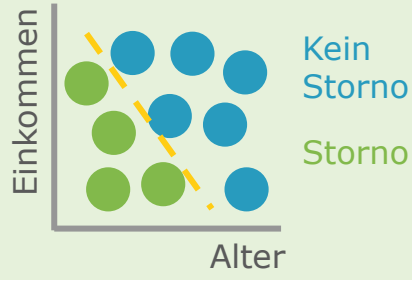
Was fangen wir damit an?

Wir beantworten Fragestellungen des überwachten Lernens…

Überwacht: Für jeden Input gibt es einen Output.

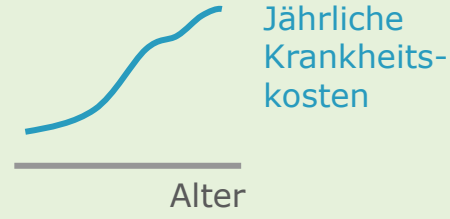
Klassifikation

- Einflussgrößen ➔ Klasse (Kategorie) der Zielgröße



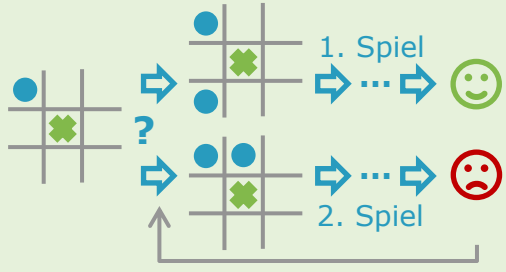
Regression

- Einflussgrößen ➔ reellwertige Ausprägung der Zielgröße



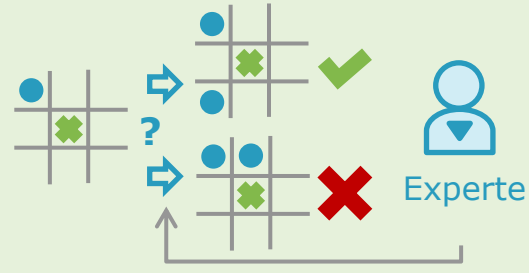
Bestärkung

- Situation ➔ Aktion ➔ Erfolg



Imitierung

- Situation ➔ Aktion ➔ beobachtete Aktion



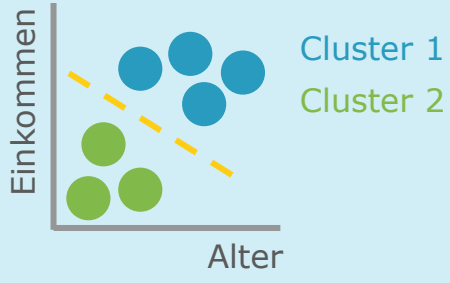
Was fangen wir damit an?

… und des unüberwachten Lernens.

Unüberwacht: Es gibt keinen klar definierten Output.

Clustering

- Merkmale ➔ Gruppen



Assoziation

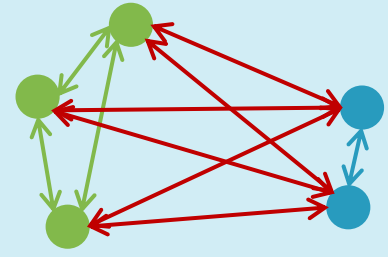
- Merkmale ➔ Kombinationen

Bier, Windeln, Milch
Bier, Windeln, Eier
Brot, Zeitung, Mehl
Bier, Windeln, Mehl
Bier, Mehl
Bier, Windeln, Milch

→ Bier + Windeln

Ähnlichkeit

- Paare ➔ Ähnlichkeitsmaß

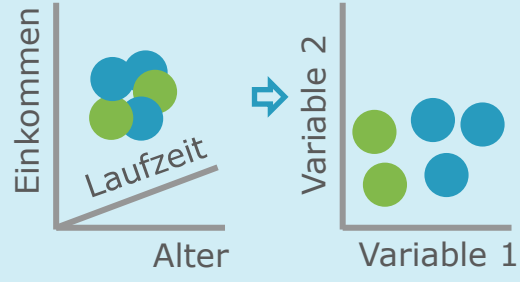


Anomalitäten

- Merkmale ➔ Ausreißer

Dimensionsreduktion

- viele Merkmale ➔ wenige Merkmale



Was fangen wir damit an?

Erfolgsfaktoren

- **Aufgabenstellung:** Was soll das Modell tun können?
  - Die Zielsetzung muss möglichst **genau konkretisiert** werden! Nur dann kann sie in eine statistische Modellanforderung **überführt** und das Modell damit **zielgerichtet** trainiert werden.
- **Performancemessung:** Wie soll das Modell bewertet werden?
  - Ein konkretes Gütemaß *im Einklang mit der Aufgabenstellung* ist von **zentraler Bedeutung**! Bei einer ungeeigneten Bewertung wird das Modell nicht das Richtige liefern.
- **Datenanforderung:** Mit welchen Daten kann und soll das Modell kalibriert werden?
  - Ein Modell kann nur **Muster** identifizieren, die **in den zugrundeliegenden Daten** vorhanden sind! Fehlende Daten bedeuten verpasstes Potenzial; unnötige Daten erschweren den Prozess.
- **Deployment:** Wie wird ein erfolgreiches Modell in die Geschäftsprozesse integriert?
  - Mit dem finalen Modell erfolgt die genaue Abschätzung der **Zielerreichung**. Schon beim Deployment muss zwingend das **zukünftige Controlling** und die **Weiterentwicklung des Modells** vorgesehen werden!

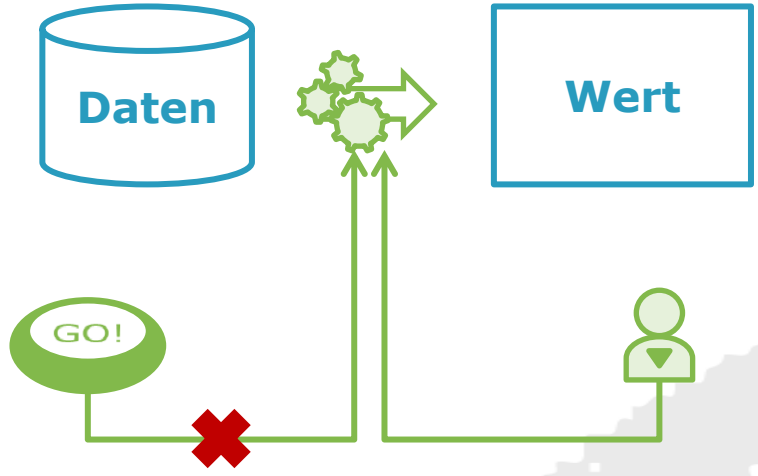
Grundvoraussetzungen für eine **Wertschöpfung** sind

- fachliche Expertise (klar definierte Ziele inkl. Maßnahmen, Datenkenntnis, …) und
- statistisches Knowhow für deren Transfer in ein modernes Data-Analytics-Modell.

Was fangen wir damit an?

Fazit

Mit Data Analytics ergibt sich als zentraler **unternehmerischer Nutzen**,
aus **Daten** einen **(Mehr-)Wert** für unternehmerische Entscheidungen zu generieren.



Dabei gilt nicht…

auf Knopfdruck
(Mehr-)Wert
generieren

…sondern:

Skills
Erfahrung
Aufwand

Grundvoraussetzungen für eine **Wertschöpfung** sind

- fachliche Expertise (klar definierte Ziele inkl. Maßnahmen, Datenkenntnis, …) und
- statistisches Knowhow für deren Transfer in ein modernes Data-Analytics-Modell.

Data Analytics
Agenda

Data Analytics: Was ist das eigentlich?

Ein Exkurs zu Bäumen und Wäldern

Was fangen wir damit an?

Institut für Finanz- und Aktuarwissenschaften

Literatur

Kontaktdaten

Beratungsangebot

Formale Hinweise

Literatur

Blome, S. und Ruß, J. (2018), Data Analytics & Co. – Was ist das eigentlich und was bringt's?, erschienen in „AssCompact" (August 2018)
https://www.ifa-ulm.de/index.php?id=177

Hahn, L. (2017), Data Analytics in der Versicherung, Vortrag auf dem Wima-Kongress 2017 der Universität Ulm (11.11. 2017)
https://www.ifa-ulm.de/index.php?id=17

Hahn, L. (2018), Machine Learning, Data Analytics und Co.: Was ist das eigentlich und viel wichtiger: Was kann man damit anfangen?, Vortrag beim Assekuranzforum LV 1/2018 in Berlin (24.04.2018)
https://www.ifa-ulm.de/index.php?id=17

Hahn, L. und Zwiesler, H.-J. (2018), Wie können Versicherer ihre Daten intelligent nutzen?, erschienen in „Versicherungswirtschaft-heute" (15.03.2018)
https://www.ifa-ulm.de/index.php?id=177

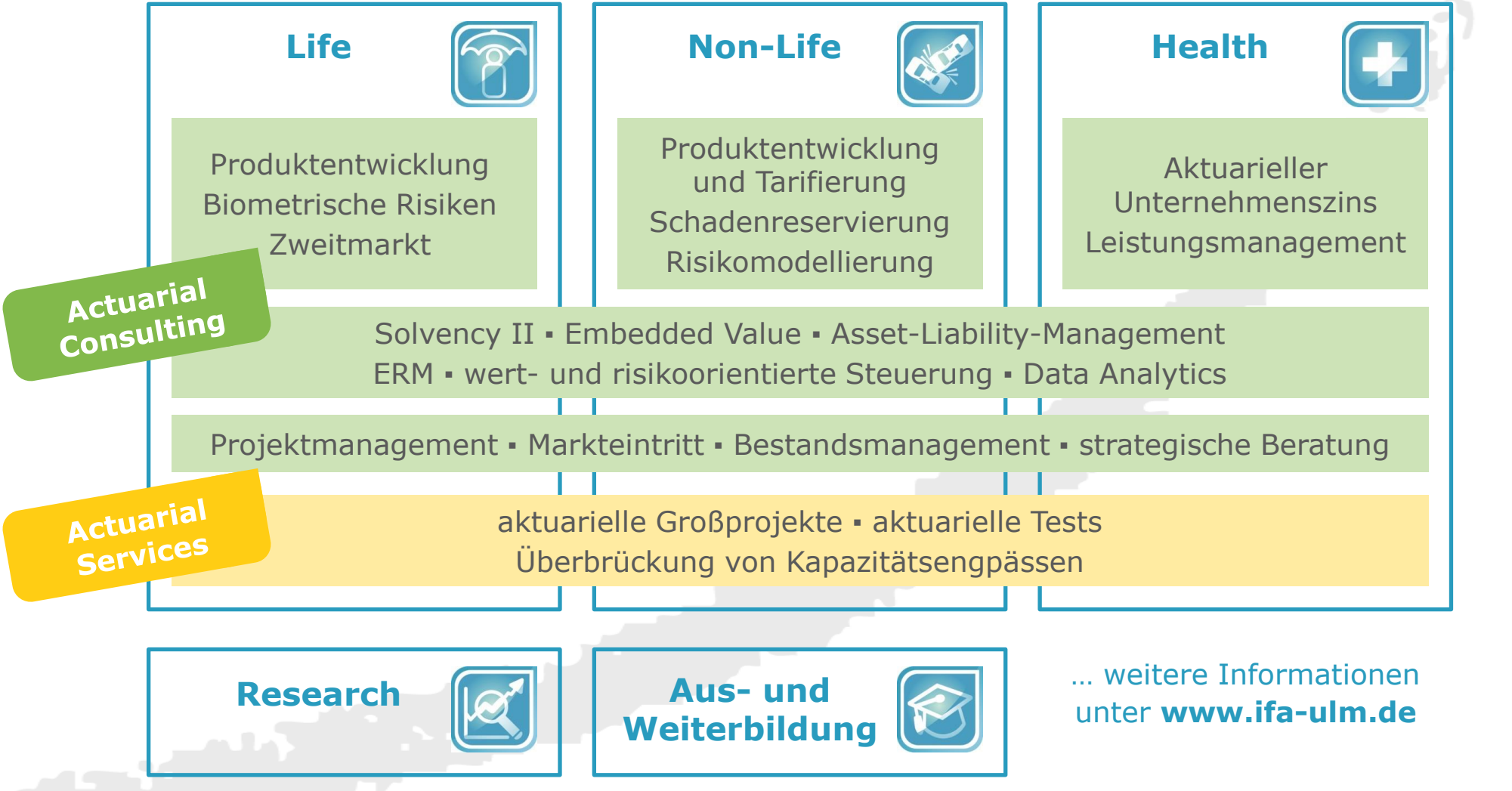
Reuß, A. (2006), Die Integration von Data-Mining in die Geschäftsprozesse von Versicherungsunternehmen – systematische Potenzialanalyse und ein generisches Prozessmodell, ifa-Verlag Ulm
https://www.ifa-ulm.de/index.php?id=239

Kontakt

Lukas Hahn
+49 731 20644-239
l.hahn@ifa-ulm.de

Institut für Finanz- und Aktuarwissenschaften

Beratungsangebot



… weitere Informationen
unter **www.ifa-ulm.de**

Institut für Finanz- und Aktuarwissenschaften

Formale Hinweise

- Dieses Dokument ist in seiner Gesamtheit zu betrachten, da die isolierte Betrachtung einzelner Abschnitte möglicherweise missverständlich sein kann. Entscheidungen sollten stets nur auf Basis schriftlicher Auskünfte gefällt werden. Es sollten grundsätzlich keine Entscheidungen auf Basis von Versionen dieses Dokuments getroffen werden, welche mit „Draft" oder „Entwurf" gekennzeichnet sind. Für Entscheidungen, welche diesen Grundsätzen nicht entsprechen, lehnen wir jede Art der Haftung ab.
- Dieses Dokument basiert auf unseren Marktanalysen und Einschätzungen. Wir haben diese Informationen vor dem Hintergrund unserer Branchenkenntnis und Erfahrung auf Konsistenz hin überprüft. Eine unabhängige Beurteilung bzgl. Vollständigkeit und Korrektheit dieser Information ist jedoch nicht erfolgt. Eine Überprüfung statistischer bzw. Marktdaten sowie mit Quellenangabe gekennzeichneter Informationen erfolgt grundsätzlich nicht. Bitte beachten Sie auch, dass dieses Dokument auf Grundlage derjenigen Informationen erstellt wurde, welche uns zum Zeitpunkt seiner Erstellung zur Verfügung standen. Entwicklungen und Unkorrektheiten, welche erst nach diesem Zeitpunkt eintreten oder offenkundig werden, können nicht berücksichtigt werden. Dies gilt insbesondere auch für Auswirkungen einer möglichen neuen Aufsichtspraxis.
- Unsere Aussagen basieren auf unserer Erfahrung als Aktuare. Soweit wir bei der Erbringung unserer Leistungen im Rahmen Ihrer Beratung Dokumente, Urkunden, Sachverhalte der Rechnungslegung oder steuerrechtliche Regelungen oder medizinische Sachverhalte auslegen müssen, wird dies mit der angemessenen Sorgfalt, die von uns als professionellen Beratern erwartet werden kann, erfolgen. Wenn Sie einen verbindlichen Rat, zum Beispiel für die richtige Auslegung von Dokumenten, Urkunden, Sachverhalten der Rechnungslegung, steuerrechtlichen Regelungen oder medizinischer Sachverhalte wünschen, sollten Sie Ihre Rechtsanwälte, Steuerberater, Wirtschaftsprüfer oder medizinische Experten konsultieren.
- Dieses Dokument wird Ihnen vereinbarungsgemäß nur für die innerbetriebliche Verwendung zur Verfügung gestellt. Die Weitergabe – auch in Auszügen – an Dritte außerhalb Ihrer Organisation sowie jede Form der Veröffentlichung bedarf unserer vorherigen schriftlichen Zustimmung. Wir übernehmen keine Verantwortung für irgendwelche Konsequenzen daraus, dass Dritte auf diese Berichte, Ratschläge, Meinungen, Schreiben oder anderen Informationen vertrauen.
- Jeglicher Verweis auf ifa in Zusammenhang mit diesem Dokument in jeglicher Veröffentlichung oder in verbaler Form bedarf unserer ausdrücklichen schriftlichen Zustimmung. Dies gilt auch für jegliche verbale Informationen oder Ratschläge von uns in Verbindung mit der Präsentation dieses Dokumentes.