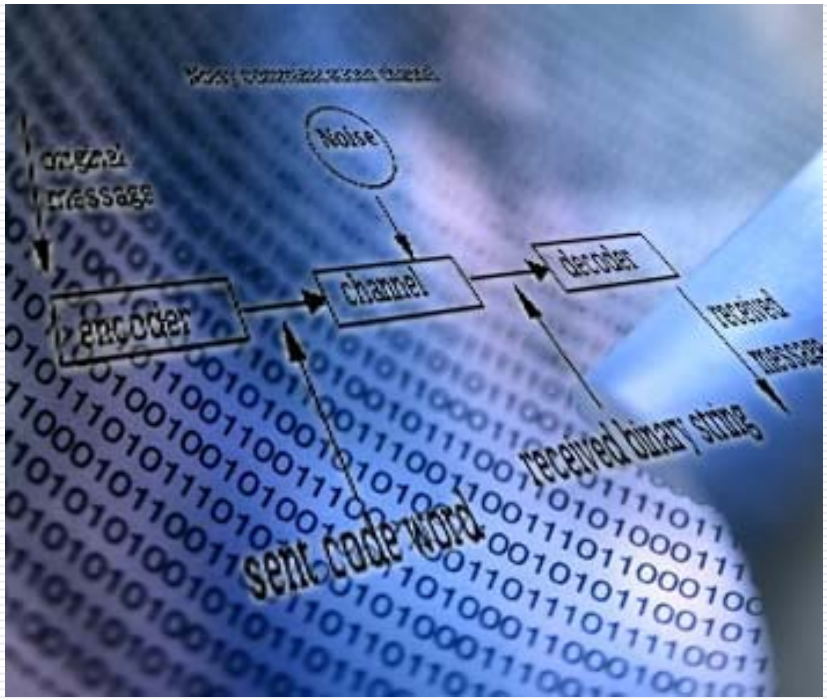
Modul 1: Einführung und Wahrscheinlichkeitsrechnung



Informationstheorie

Dozent: Prof. Dr. M. Gross

E-mail: grossm@inf.ethz.ch

Assistenten: Bernd Bickel, Daniel Cotting, Michael Waschbüsch, Boris Köpf, Andrea Francke, Ueli Peter, Barbara Scheuner

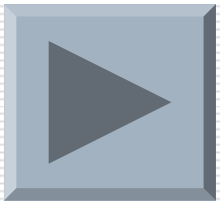
Web Page: http://graphics.ethz.ch

Allgemeines

- Zur Vorlesung
- Skript und Textbooks
- Elektronisches Material
- Tafel – Beispiele
- Übungsablauf - Gruppeneinteilung
- Testatbedingung: 8 aus 9 Übungen (Empfehlung)
- Klausur: 2 Stunden
  Hilfsmittel: keine
- Kein Midterm

Java-Applets

- Java-Applets zur Illustration der wichtigsten Algorithmen



Vorlesungsplan

Thema	Vorlesung	Uebung		
		Typ	Ausgabe	Abgabe
Einführung und Grundlagen	30.10.2006	Theorie	30.10.2006	06.11.2006
Stochastische Prozesse	06.11.2006	Theorie	06.11.2006	13.11.2006
Entropie	13.11.2006	Praxis	13.11.2006	20.11.2006
Bedingte Entropie	20.11.2006	Theorie	20.11.2006	27.11.2006
Informationsquellen	27.11.2006	Theorie	27.11.2005	04.12.2006
Codierung diskreter Quellen	04.12.2006	Keine		
Optimalcodierung und Huffman Codes	11.12.2006	Theorie	11.12.2006	18.12.2006
Arithmetische Codierung/Intervalllängen/ LZ 1. Teil	18.12.2006	Keine		
Arithmetische Codierung/Intervalllängen/ LZ 2. Teil	08.01.2007	Theorie	08.01.2007	15.01.2007
Binäre Kanäle	15.01.2007	Theorie	15.01.2007	22.01.2007
Codierungstheorem und Fehlerkorrektur	22.01.2007	Praxis*	22.01.2007	29.01.2007
Syndromcodierung/ Hamming Codes Polynomevaluationscodes	29.01.2007	Keine		

Skript und Bücher

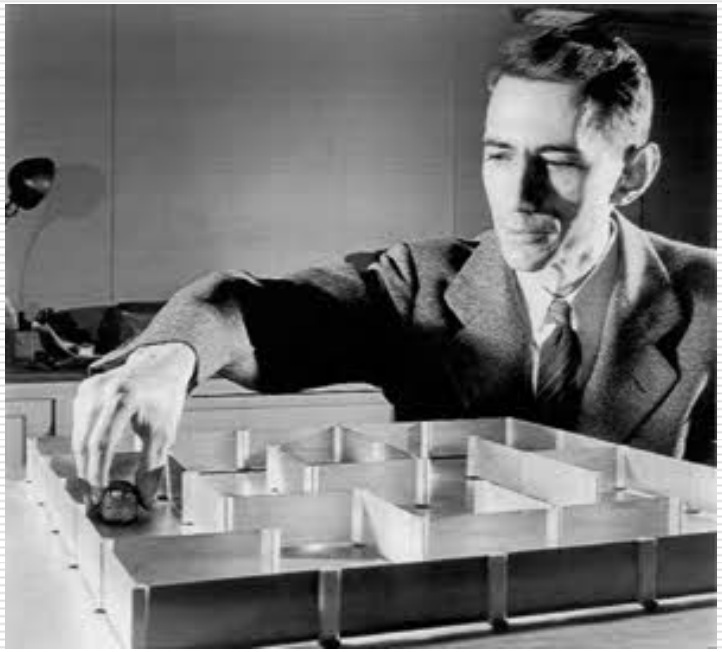
- H. Klimant, R. Piotraschke, D. Schönfeld: *Informations- und Kommunikationstheorie*, 2.Auflage, Teubner, 2003.
- T. Cover, J. Thomas: *Elements of Information Theory*, John Wiley, 1991.
- U. Maurer: *Skript zur Vorlesung Information und Kommunikation*, WS 2003/2004.
- F. Reza: *An Introduction to Information Theory*, Dover Publications, 1994.
- H.D. Lüke: *Signalübertragung*, Springer, 6. Auflage, 1995.
- T. Bell, J. Cleary, I. Witten: *Text Compression*, Prentice Hall, 1990.
- A. Oppenheim, R. Schafer, J. Buck: *Zeitdiskrete Signalverarbeitung*, 2. Auflage Pearson, 2004.

Ziele der Vorlesung

- Einführung in die Informations- und Kodierungstheorie
- Quantifizierung von Information
- Abschätzung mathematischer Grenzen für die Kompression von Daten
- Verlustfreie Kodierungsverfahren
- Redundante, fehlerkorrigierende Kodierungsverfahren
- Praktische Beispiele

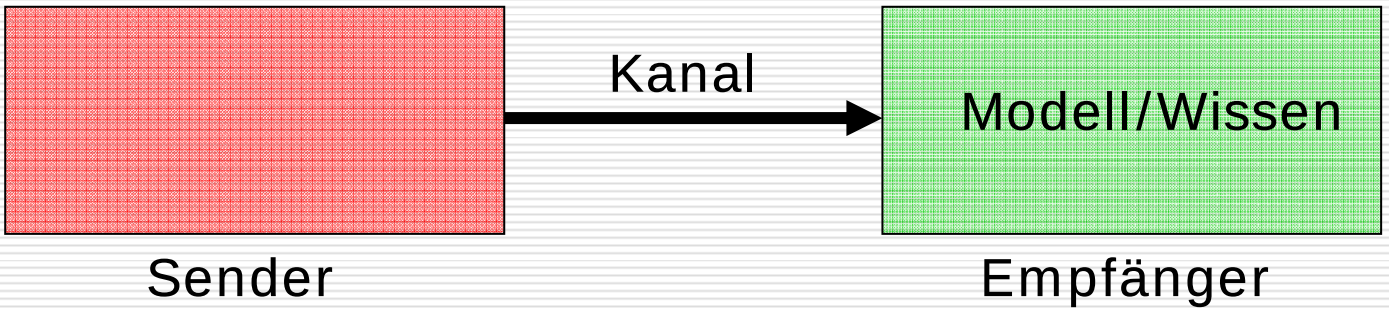
Historie

- Informationstheorie wurde von Claude Shannon begründet und 1948 publiziert
- “Mathematical theory of communication”, eine der bedeutendsten Theorien der Informatik



Begriff der Information

- Information bekommen wir, wenn wir etwas „Neues“ erfahren
- Altbekanntes stellt keine Information dar
- Information misst also den Neuheitsgrad einer empfangenen Meldung



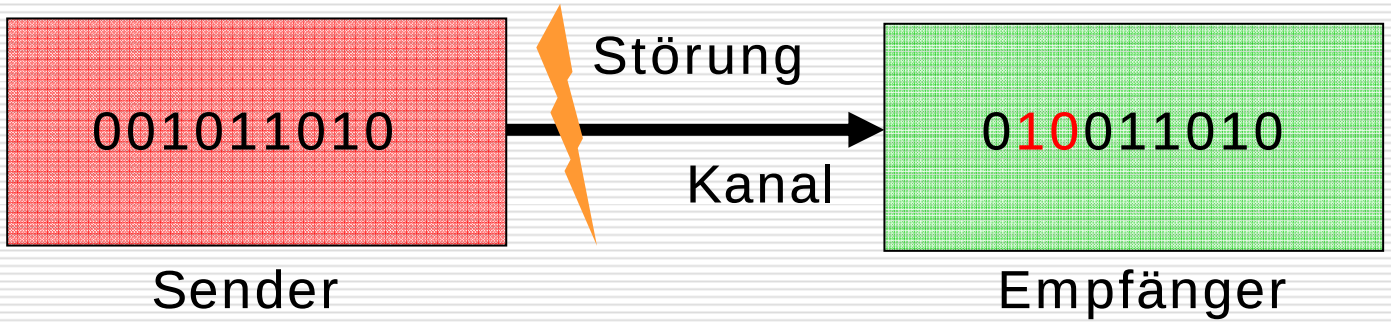
Information ist also vom Kenntnisstand des Empfängers abhängig.

Kommunikation

- Der Empfänger erhält dann u. U. verfälschte Information
- Fehler können durch Einfügen von Redundanz gezielt korrigiert werden
- „Weiss“ der Empfänger mehr, so muss der Sender weniger Information übertragen
- Offenbar ist der Begriff der Information eng mit den Begriff der Kommunikation verknüpft
- Kommunikation ist der Austausch von Information zwischen zwei oder mehreren Partnern

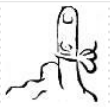
Übertragungsmodell

- Der Sender stellt eine Informationsquelle dar
- Diese kann kontinuierlich oder diskret sein
- Der Kanal kann ideal (verlustfrei), oder nicht-ideal (verlustbehaftet) sein
- In nicht-idealen Kanälen entstehen durch Störungen Fehler



Kompression

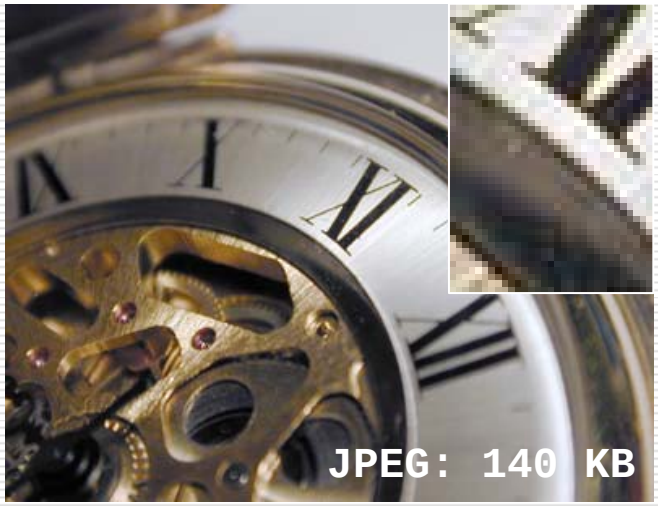
- Ziel ist die Übertragung von Information mit möglichst wenig Informationseinheiten
  - Diese werden auch als bits (basic information units) bezeichnet.
- Verhältnis aus überschüssigen bits zu notwendigen bits ist die Redundanz
- Kompression zielt auf die Entfernung unnötiger, redundanter bits
- Kompression kann verlustfrei oder verlustbehaftet sein



bits werden grundsätzlich von Bits (binary digits) unterschieden.
Bei Binärcodierung gilt: bits = Bits

Beispiele

- Verlustbehaftete Bildkompression:



- Verlustlose Textkompression:
  „Dies ist ein kleiner Text“
  25 ASCII-Zeichen = 200 Bits

Grundsätzliche Fragen

- Gibt es eine untere Grenze für die minimale Anzahl von bits zur Übertragung einer bestimmten Information?
- Können wir den „Informationsgehalt“ einer Nachricht quantitativ erfassen?
- Wie ändert sich die Betrachtung, wenn wir die Verhältnisse im statistischen Mittel über einen langen Zeitraum betrachten?
- Gibt es Möglichkeiten, diese theoretischen unteren Schranken zu erreichen?

Beispiel: Würfeln

- Würfelexperiment mit 6 gleichwahrscheinlichen Ereignissen – diskret, binärcodiert
- Um ein Ereignis zu codieren, brauchen wir offenbar

$$\lceil \log_2 6 \rceil = 3 \,\text{Bits}$$

- Für *k* Würfe benötigen wir demnach

$$\lceil k \log_2 6 \rceil \,\text{Bits}$$

- Das heisst, 3 Würfe ($6^3 = 216$) können wir mit 8 Bits codieren
- Für $k \to \infty$ erhalten wir 2.585 Bits pro Ereignis

Informationsgehalt

- Um eine Zufallsvariable mit N verschiedenen, gleichwahrscheinlichen Zuständen binär zu codieren, benötigen wir offenbar

$$\lceil \log_2 N \rceil \text{Bits}$$

- Sei $p_N = 1/N$ die Wahrscheinlichkeit eines Zustandes, so benötigen wir also

$$\lceil -\log_2 p_N \rceil \text{Bits}$$

- Die Zustandswahrscheinlichkeit spielt also bei der Codierung eine bedeutende Rolle

Grenzbetrachtung

- Wir verallgemeinern dieses Konzept
- Sei Z eine Zufallsvariable mit N möglichen Zuständen $\{z_1,\ldots,z_N\}$
- Sei p_i die Wahrscheinlichkeit, dass $Z=z_i$, so könnte man mit folgender Verallgemeinerung die Anzahl der benötigten Bits berechnen:

$$-\sum_{i=1}^{N} p_i \log_2 p_i$$

- Erklärung folgt später

Man denke über die Implikationen dieser Formel gut nach.

Beispiel: Textcodierung

Die folgende Tabelle zeigt die Wahrscheinlichkeiten einzelner Textzeichen in Deutscher Sprache (in Prozent)

a	b	c	d	e	f	g	h	i	j	k	l	m
6.44	1.93	2.68	4.83	17.5	1.65	3.06	4.23	7.73	0.27	1.46	3.49	2.58
n	o	p	q	r	s	t	u	v	w	x	y	z
9.84	2.9	0.96	0.02	7.54	6.83	6.13	4.17	0.94	1.48	0.04	0.08	1.14

Beispiel: Textcodierung

- Mit Hilfe der vorherigen Formel berechnen wir die Anzahl von Bits zu Codierung eines einzelnen Zeichens:

$$-\sum_{i=1}^{N} p_i \log_2 p_i = 4.07$$

- Diese Grösse wird auch Entropie genannt.
- Sie stellt einen statistischen Mittelwert dar, d.h. im Mittel braucht man mindestens 4.07 Bits zur Codierung eines Zeichens in Deutscher Sprache.

Die Wahrscheinlichkeit

- Ein Wahrscheinlichkeitsmass auf einer Menge Ω ist eine Funktion P von Untermengen von Ω auf $\Re$, welche die folgenden Axiome erfüllt:
  - $P(\Omega) = 1$
  - Wenn $A \subset \Omega$, dann $P(A) \geq 0$
  - Wenn A_1 und A_2 disjunkt, dann

$$P(A_1 \cup A_2) = P(A_1) + P(A_2)$$

  - Allgemein (Summenformel):

$$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$$

Eigenschaften

- Additionsgesetz

$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$$

Beispiel: 2 Münzwürfe

- Praktische Berechnung von Wahrscheinlichkeiten durch Zählen:

$$\mathrm{P}(A) = \frac{Anzahl\ Ereignisse\ mit\ A}{Gesamtzahl\ Ereignisse}$$

- P(*A*): Kopf im ersten Wurf
- P(*B*): Kopf im zweiten Wurf

$$\Omega = \{kk, kz, zk, zz\}$$

- P(*C*): Kopf im ersten oder zweiten Wurf

$$\begin{aligned}
&\mathrm{P}(C) \neq \mathrm{P}(A) + \mathrm{P}(B), \\
&\mathrm{P}(C) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B) \\
&\qquad = 0.5 + 0.5 - 0.25 \\
&\qquad = 0.75
\end{aligned}$$

Bedingte Wahrscheinlichkeit

- Die bedingte Wahrscheinlichkeit ist die Wahrscheinlichkeit von *A* gegeben *B*:

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}$$

- Daraus folgt das Multiplikationsgesetz

$$\mathrm{P}(A \cap B) = \mathrm{P}(A|B)\,\mathrm{P}(B)$$

Beispiel: Urne (1)

- Urne mit 3 roten und einem blauen Ball.
  Zwei mal ziehen ohne zurücklegen.
  - R_1 : rot im ersten Zug
  - R_2 : rot im zweiten Zug

$$\begin{aligned} \mathrm{P}(R_1 \cap R_2) &= \mathrm{P}(R_1) \cdot \mathrm{P}(R_2 \mid R_1) \\ &= \frac{3}{4} \cdot \frac{2}{3} \\ &= \frac{1}{2} \end{aligned}$$

Totale Wahrscheinlichkeit

- Seien $B_1,\ldots,B_n$ so, dass

$$\bigcup_{i=1}^{n} B_i = \Omega, \text{ und } B_i \cap B_j = \emptyset, i \neq j$$

- Sowie $P(B_i) > 0$ für alle i. Dann gilt für ein beliebiges Ereignis A

$$P(A) = \sum_{i=1}^{n} P(A|B_i)\,P(B_i)$$

Beispiel: Urne (2)

- Wahrscheinlichkeit, rot im zweiten Zug zu ziehen.

$$
\begin{aligned}
\mathrm{P}(R_2) &= \mathrm{P}(R_2 \mid R_1) \cdot \mathrm{P}(R_1) + \mathrm{P}(R_2 \mid \overline{R}_1) \cdot \mathrm{P}(\overline{R}_1) \\
&= \frac{2}{3} \cdot \frac{3}{4} + 1 \cdot \frac{1}{4} \\
&= \frac{3}{4}
\end{aligned}
$$

Bayessche Regel

- Die Bayessche Regel ist von fundamentaler Bedeutung in der Wahrscheinlichkeitstheorie
- Seien $B_1,\ldots,B_n$ Ereignisse so, dass

$$\bigcup_{i=1}^{n} B_i = \Omega, \text{ und } B_i \cap B_j = \emptyset, i \neq j$$

- Sowie $P(B_i) > 0$ für alle i. Dann gilt:

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)} \qquad P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

Beispiel: Bayes (1)

- Spam-Filter
  - + : Keyword in Mail
  - – : Keyword nicht in Mail
  - *S* : Mail ist Spam
  - *N* : Mail ist kein Spam
- Statistische Auswertungen ergeben:
  - $\mathrm{P}(+ \mid S) = 0.88$
  - $\mathrm{P}(- \mid S) = 0.12$
  - $\mathrm{P}(+ \mid N) = 0.14$
  - $\mathrm{P}(- \mid N) = 0.86$

Beispiel: Bayes (2)

- Formel von Bayes angewandt:

$$P(N|+) = \frac{P(+|N) \cdot P(N)}{P(+|N) \cdot P(N) + P(+|S) \cdot P(S)}$$

- P(*N*) = 0.5 und P(*S*) = 0.5

$$P(N|+) = \frac{0.14 \cdot 0.5}{0.14 \cdot 0.5 + 0.88 \cdot 0.5} \approx 0.13$$

Zufallsvariablen

- Eine Zufallsvariable *X* ist im Wesentlichen eine Zufallszahl
- Sie kann entweder kontinuierlich oder diskret sein
- Diskrete Zufallsvariablen nehmen nur endlich viele, oder unendlich viele, aber abzählbare Zustände an!

- Beispiel: Der Würfelwurf als Zufallsvariable *X* mit Werten 1,2,3,4,5,6

Zufallsvariablen

- Die Wahrscheinlichkeit auf einer (diskreten) Zufallsvariable wird wie folgt definiert:

  Seien x_1, x_2... die möglichen Werte von X, dann ist die Funktion $p(x_i)=P(X=x_i)$ die Häufigkeitsfunktion (frequency function)

- Es gilt: $$\sum_{i=1}^{n} p(x_i) = 1$$

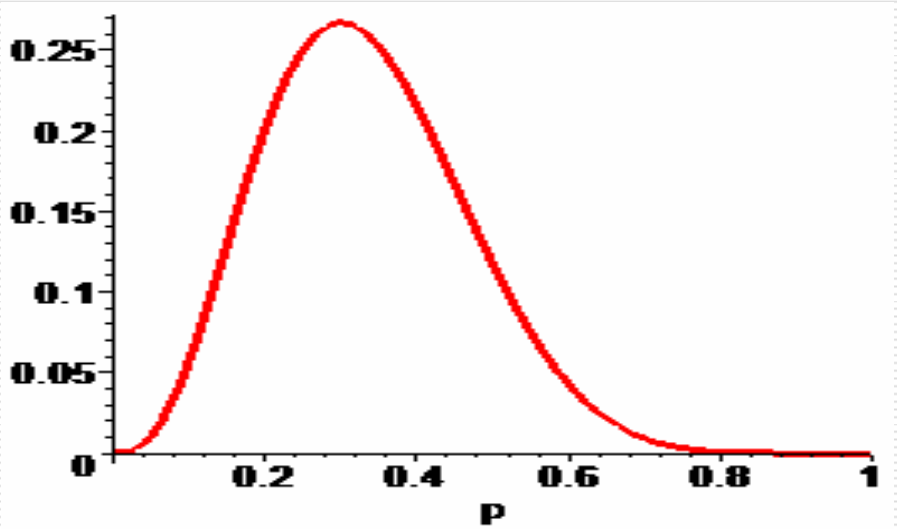
- Zwei Zufallsvariablen X und Y sind unabhängig, wenn

$$P(X = x_i, Y = y_j) = P(X = x_i)\,P(Y = y_j)$$

Verteilungen

- Zufallsvariablen sind oft charakteristisch verteilt
- Eine bekannte Funktion ist die Binomialverteilung

$$\binom{n}{k} p^k (1-p)^{n-k}$$



Verbundverteilungen

- Ebenso kann man die gemeinsame Wahrscheinlichkeitsverteilung mehrerer Zufallsvariablen untersuchen
- Die Verbundwahrscheinlichkeit von *X* und *Y*

$$\mathrm{p}(x_i, y_j) = \mathrm{P}(X = x_i, Y = y_j)$$

oder

$$\mathrm{p}(x_1 \cdots x_n) = \prod_{i=1}^{n} \mathrm{P}(X = x_i)$$

- Notation:

$$\mathrm{p}_X(x_i) = \mathrm{P}(X = x_i)$$

Marginalisierung

- Die Wahrscheinlichkeit eines Zustandes *x* von *X* kann durch Marginalisierung (Summation) aus Verbundwahrscheinlichkeit berechnet werden

$$p_X(x) = \sum_i p(x, y_i)$$

- Für *m* Zufallsvariablen gilt entsprechend

$$\mathrm{p}_{X_1}(x_1) = \sum_{x_2 \ldots x_m} \mathrm{p}(x_1 \cdots x_m)$$

$$\mathrm{p}_{X_1 X_2}(x_1, x_2) = \sum_{x_3 \ldots x_m} \mathrm{p}(x_1 \cdots x_m)$$

Bedingte Verteilungen

- Die bedingte Verteilung von *X* und *Y*
- Die bedingte Wahrscheinlichkeit von *X* und *Y* ist

$$\mathrm{P}(X = x_i \,|\, Y = y_j) = \frac{\mathrm{P}(X = x_i, Y = y_j)}{\mathrm{P}(Y = y_j)} = \frac{\mathrm{p}_{XY}(x_i, y_j)}{\mathrm{p}_Y(y_j)}$$

- Vergleiche mit Bayesschem Gesetz!

- Oder auch $\mathrm{p}_{XY}(x, y) = \mathrm{p}_{X|Y}(x|y)\,\mathrm{p}_Y(y)$

- Marginalisierung

$$\mathrm{p}_X(x) = \sum_y \mathrm{p}_{X|Y}(x|y)\,\mathrm{p}_Y(y)$$

Beispiel: Bed. Verteilungen

- Gegeben X und Y mit Verteilungen:

$x \backslash y$	0	1	2	3
0	1/8	2/8	1/8	0
1	0	1/8	2/8	1/8

  - Marginalisierung: $P_Y(1) = P_{XY}(0,1) + P_{XY}(1,1)$
$= \frac{2}{8} + \frac{1}{8} = \frac{3}{8}$

  - $P_{X|Y}(0|1) = \frac{P_{XY}(0,1)}{P_Y(1)} = \frac{2/8}{3/8} = \frac{2}{3}$

  - $P_{X|Y}(1|1) = \frac{P_{XY}(1,1)}{P_Y(1)} = \frac{1/8}{3/8} = \frac{1}{3}$