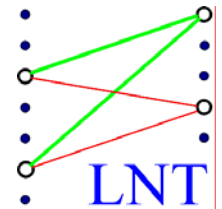


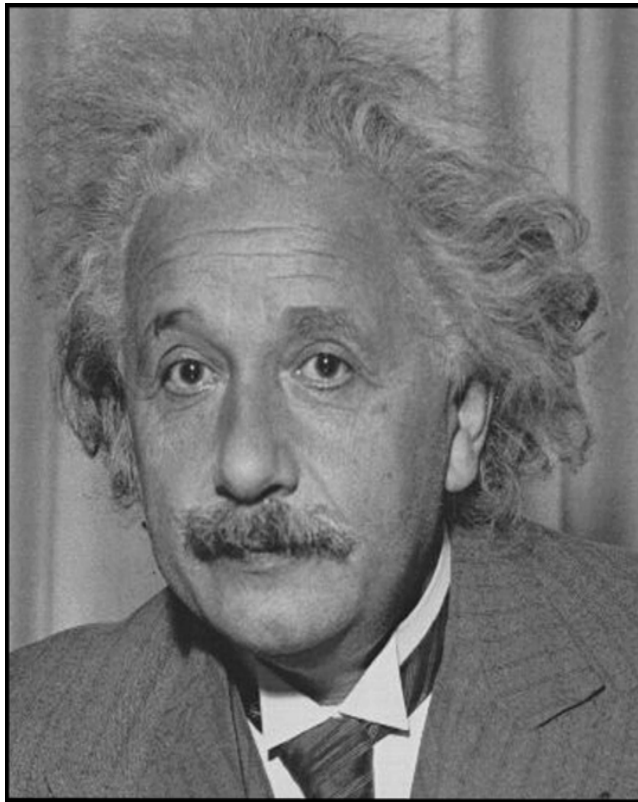
Informationstheorie: Eine praktische Wissenschaft

Prof. Dr. Ing. Joachim Hagenauer
Lehrstuhl für Nachrichtentechnik
Technische Universität München
Bayerische Akademie der Wissenschaften

© 2004 J. Hagenauer

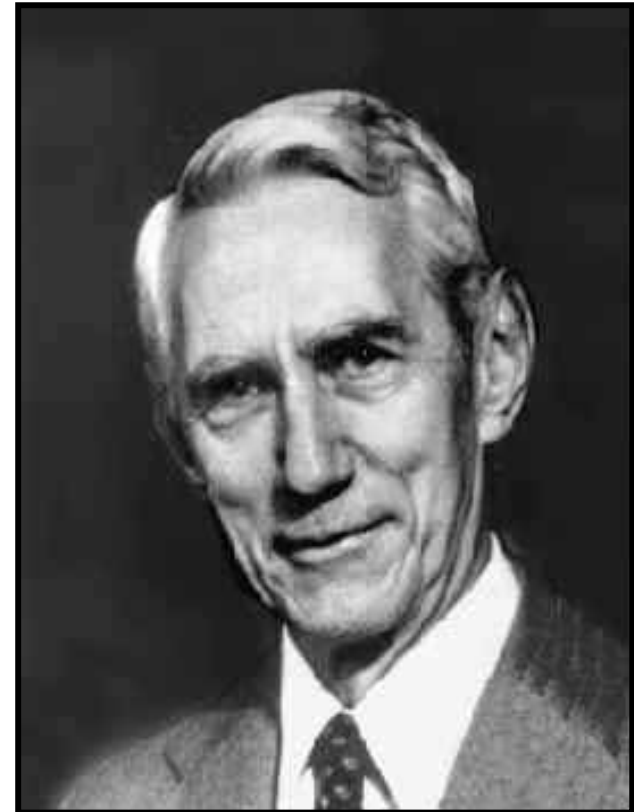


Kennen Sie Einstein?



$$E = m \cdot c^2$$

Kennen Sie Shannon?



$$C = \frac{1}{2} \log\left(1 + \frac{2E_s}{N_0}\right)$$

Lebenslauf: Claude Elwood Shannon

- Geboren am 30. April 1916
- 1932 BSc in Elektrotechnik und Mathematik an der Univ. of Michigan
- 1936 MSc Arbeit am MIT: "Boolean Algebra is Useful in Circuit Design"
- 1943 PhD Thesis: "An Algebra for Theoretical Genetics"
- Zweimal in der Deutsch-Prüfung durchgefallen
- 1941 - 1956 Bell Laboratorien
- 1948 Begründung der Informationstheorie
- 1956 - 1966 Professor am MIT
- Weitere Erfindungen: Schachcomputer, Balljongliermaschine, Maus im Labyrinth, etc.
- Gestorben am 27. Februar 2001

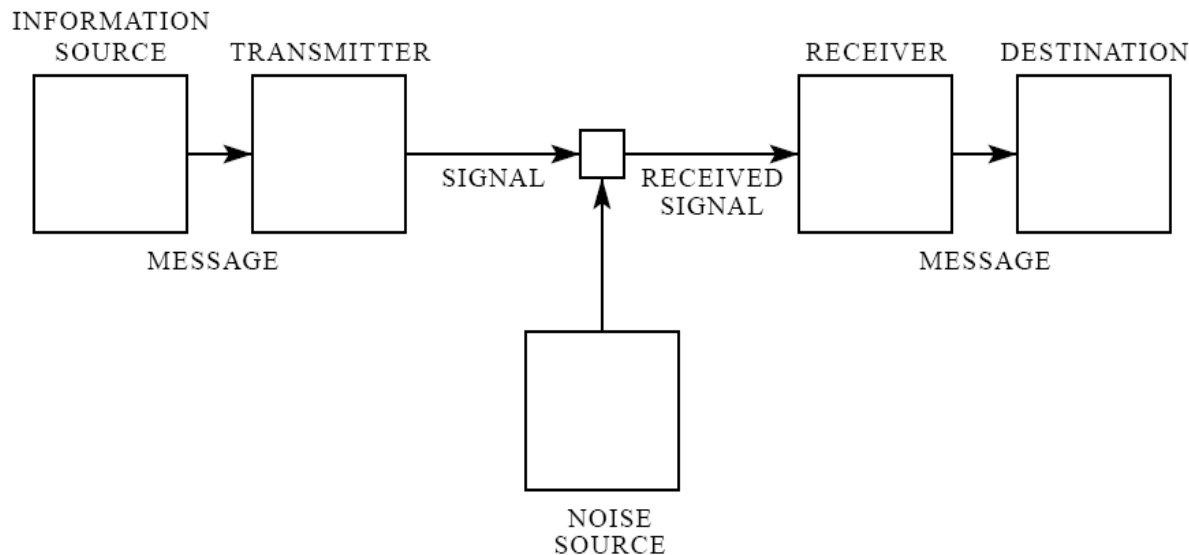
Claude Elwood Shannons grundlegende Veröffentlichungen zur Informationstheorie:

- C.E. Shannon, "A Mathematical Theory of Communication", Bell System Technical Journal, Vol. 27 (July and October 1948), pp. 379-423 and 623-656.
- C.E. Shannon, "Communication Theory of Secrecy Systems", Bell System Technical Journal, Vol. 28 (1949), pp. 656-715.
- C.E. Shannon, "Prediction and Entropy of Printed English", Bell System Technical Journal, Vol. 30 (1951), pp. 50-64.

Umberto Eco, Einführung in die Semiotik, 7. Ausg., München, Fink 1991, Uni-Taschenbücher, S. 47:

„Wenn jedes Kulturphänomen ein Kommunikationsphänomen ist, dann muss man die elementare Struktur der Kommunikation dort aufsuchen, wo Kommunikation sozusagen minimal stattfindet, d.h. auf der Ebene der Übertragung von Information zwischen zwei Apparaten.“

Shannons Modell der Informationsübertragung



Was ist Information?

Information ist die **Verringerung von Unsicherheit**

Wie misst man Unsicherheit?

Ein Ereignis trete mit Wahrscheinlichkeit p_i auf.

Es hat dann einen Informationsgehalt von

$$\log_2 \frac{1}{p_i}$$

Shannons Maß für die **mittlere Unsicherheit** ist der mittlere Informationsgehalt (die **Entropie**),

gemessen in **binary digit (bit)**

$$H = \sum_{i=1}^N p_i \log_2 \frac{1}{p_i} \text{ bit}$$

Beispiele für den Informationsgehalt von statistischen Ereignissen

- Münze

$$H = \sum_{i=1}^2 \frac{1}{2} \log_2 \frac{1}{1/2} \text{ bit} = \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 1 = 1 \text{ bit}$$

- Würfel $H = 2,6 \text{ bit}$

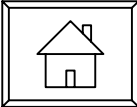
Differenz $H = 1,4 \text{ bit}$

- gezinkter Würfel $H = 1,2 \text{ bit}$

- Lotto (6 aus 49) $H = 23,7 \text{ bit}$

Was ist der Informationsgehalt (Entropie) von Texten?

Entropie von Buchstabengruppen:

- **Textanalyse:** Häufigkeit von Buchstabengruppen in der Bibel
- **Textsynthese:** Statistische Texterzeugung aus der Häufigkeit von Buchstabengruppen
- **Demonstration:** Bibeltext 

Teil des synthetisch erzeugten Bibeltexts

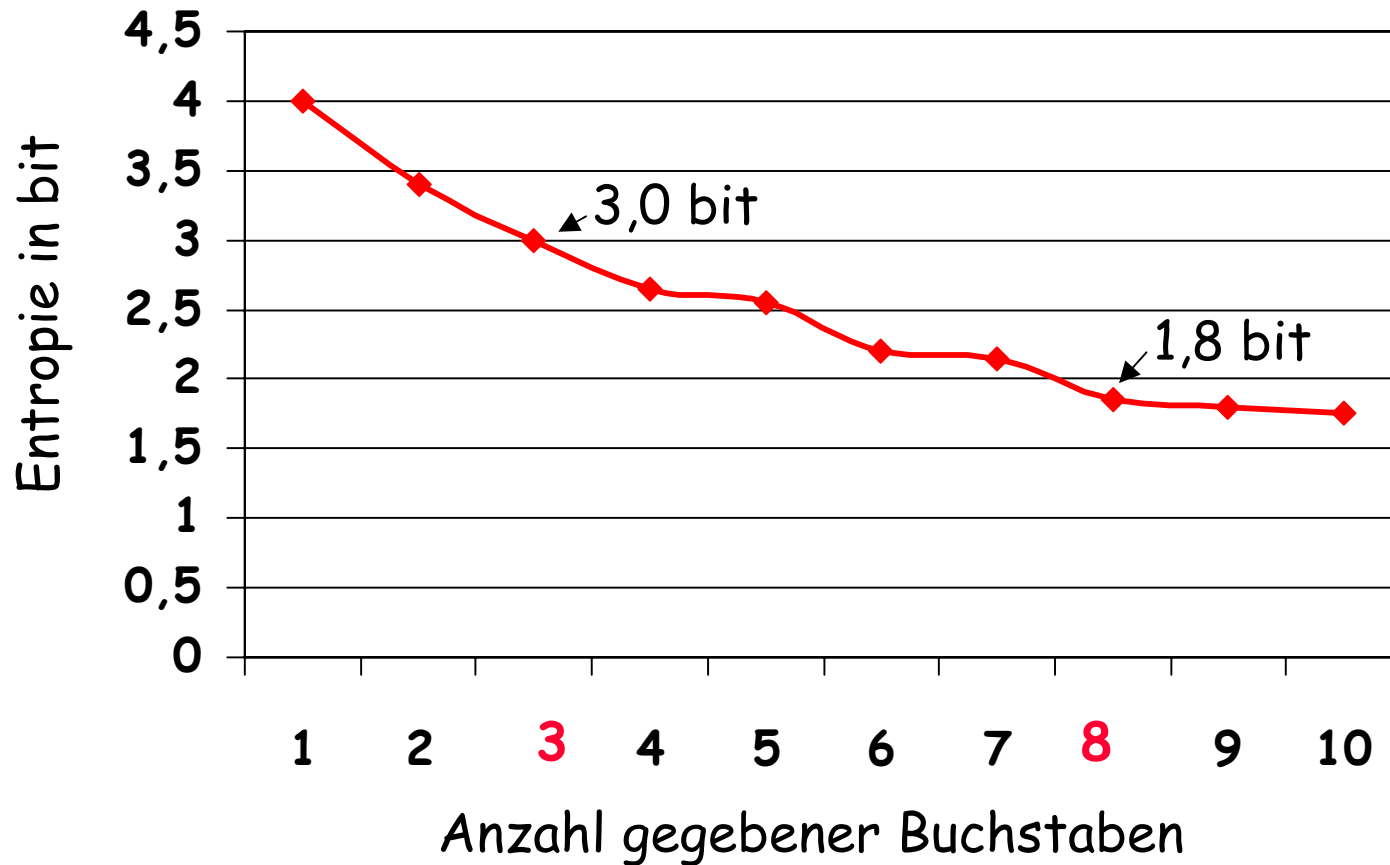
gottageberbocht mit aus tagen kinechiner acht
trachasseja die manden leneher sieseinem res zu blanks
leade david wes war ermen dassainen letwollsbaum
zwarach wie dasserwein der und bricher gefanich sach
bewar mich fer begilebordet mit hatte vere
knaalbsunglastagen stetwarubt st ammen dass einzenn den
den wa du sicklenn ihm dies tres herd den wer gend jiskne
kobarblut volkste sprachtenst inete sehrt judarberinand
sollebophockte braussephas wangeges seist deiblucht get
ben wille we den meinamte ger mosen sich sie der kommt

Was ist der ultimative Informationsgehalt (Entropie) von Texten?

Shannons Experiment: (wiederholt von Küpfmüller)
Was ist Ihre mittlere Unsicherheit beim Raten des
nächsten Buchstabens?

Der Vortrag von Prof. Regine Kahmann, korrespondierendes Mitglied der Bayerischen Akademie der Wissenschaften, geht am 29. November 2004 der Frage nach, wie man Pflanzen züchten kann, die resistent gegen Pilzbefall sind. Man weiß inzwischen, dass Resistenz gegen Pilze angeboren ist und durch sogenannte Resistenzgene vermittelt wird. Wie aber gelingt es Pilzen, eine Pflanze erfolgreich zu besiedeln und sich dort zu vermehren?

Was ist der ultimative Informationsgehalt (Entropie) von Texten?



Die mittlere Unsicherheit (Entropie) ist 1.5 bit/Buchstabe

Shannons Kompressionstheorem:

Zur Darstellung eines Textes benötigt man bei einem Alphabet von $32 = 2^5$ Buchstaben nur

$$H = 1.5 \text{ bit/Buchstabe}$$

statt

5 bit/Buchstabe

Textdateien im PC können also etwa um den Faktor 3 komprimiert werden. (z.B. mit WinZip)

Kompressionsgewinn in Abhängigkeit von der Dateigröße bei Kompressionsverfahren nach Lempel&Ziv LZ77

Buchstaben	Gewinn %
100	2%
400	15%
1.000	23%
4.000	40%
10.000	47%
40.000	53%
100.000	56%

Praktische Folgerung:

Komprimieren Sie immer ein Buch nicht nur ein Kapitel !

Andere Kompressionsverfahren:

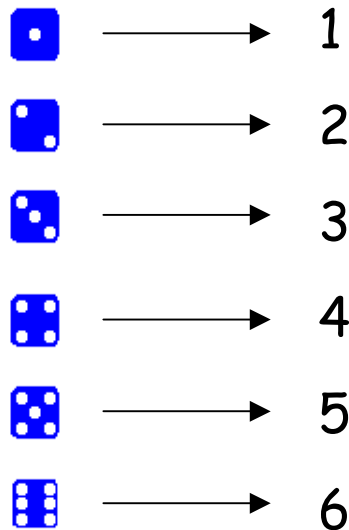
- Huffman-Codes 1950 (z.B im FAX und in MP3)
- Context-Tree-Weighting Algorithmus, Willems 1995 (erreicht bei großen Texten 1,65 bit/Buchstabe)
- DNACompress, Chen 2002, angepasst auf den genetischen Code

Shannons Theorie der Nachrichtenübertragung:

Was kommt rüber?

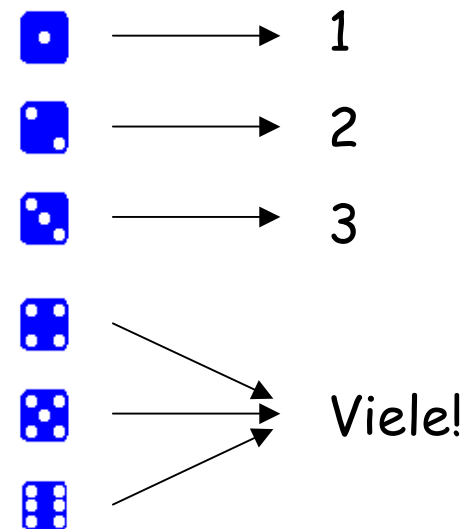
Unsicherheit vorher minus Unsicherheit hinterher
= übertragene Information

Perfekte Übertragung



$$2,6 \text{ bit} - 0 \text{ bit} = 2,6 \text{ bit}$$

Fehlerhafte Übertragung



$$2,6 \text{ bit} - 0,8 \text{ bit} = 1,8 \text{ bit}$$

Shannons Theorie der Nachrichtenübertragung:

Shannon gibt für jeden Übertragungskanal (z.B. Funkkanal) eine Kapazität gemessen in bit pro Kanalbenutzung an.

Diese hängt ab von

- der für jedes Zeichen empfangenen Energie E_s
- der Störung N_0

$$C = \frac{1}{2} \log_2 \left(1 + \frac{2E_s}{N_0} \right)$$

Shannons Theorie der Nachrichtenübertragung:

Die in jedem Empfänger enthaltene thermische Störung wird statistisch nach Gauß beschrieben und berechnet sich aus Boltzmann-Konstante mal Temperatur des Empfängers

$$N_0 = k_B \cdot T_K$$



K. F. Gauß 1777-1855



Gauß Statistik

Shannons Theorie der Nachrichtenübertragung:

Wieviel Energie braucht man
um eine SMS (max. 1000 bits) zum Handy zu übertragen?

Die Kanalkapazität C wird nur zur Hälfte genutzt, da man
Redundanz mit überträgt

$$C = \frac{1}{2} \log_2 \left(1 + \frac{2E_s}{N_0} \right) = \frac{1}{2}$$

Dies führt auf eine benötigte Empfangsenergie pro bit von

$$E_s = \frac{1}{2} N_0 = \frac{1}{2} k_B T_K = \frac{1}{2} 1,38 \cdot 10^{-23} \frac{Ws}{K} \cdot 300 \text{ K} \approx 2 \cdot 10^{-21} Ws$$

Die Sendeenergie muss wegen der starken Dämpfung um den
Faktor $10^{11} = 100.000.000.000$ höher sein

Der Sender benötigt also pro SMS ca. 2×10^{-7} Ws (Joule)
Von dieser theoretischen Shannongrenze sind wir derzeit
etwa um den Faktor 100 entfernt!

Anwendung der Shannonschen Informationstheorie

am Lehrstuhl für Nachrichtentechnik der TUM (LNT, P. Hanus, B. Göbel)

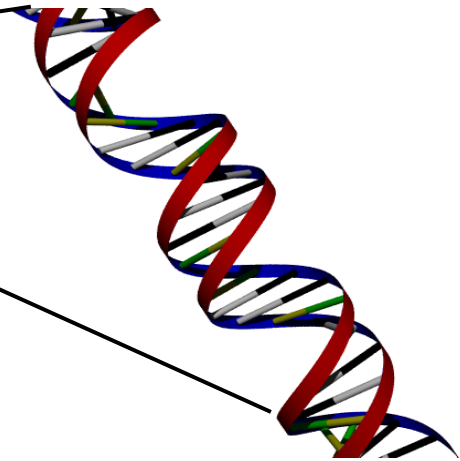
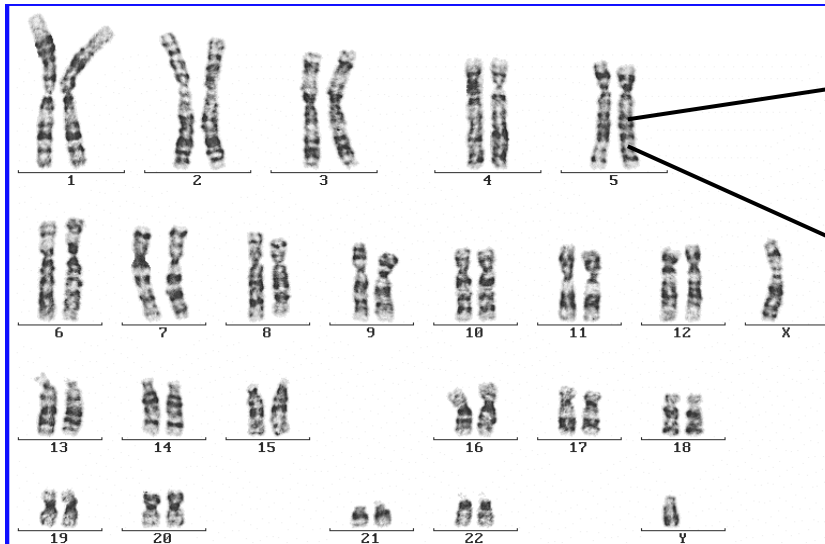
- Informationsübertragung von genetischen Abweichungen zu Krankheiten wie Schizophrenie und Parkinson (mit Dr. Müller, GSF, Inst. f. Humangenetik Neuherberg)
- Die Klassifizierung von genetischen DNS-Sequenzen, um entwicklungsgeschichtliche Bäume aufzustellen
- Die Inhaltserkennung, d.h. die Erkennung von Texten mit unbekannter Autorenschaft

Anwendung der Shannonschen Informationstheorie auf den genetischen Code

DNS: Alleiniger Träger der genetischen Information

- Doppel-Helix -> 2 komplementäre Stränge von vier Basen (G,A,T,C)
- 46 Chromosome, 23 von jedem Elternteil
- 3×10^9 Basenpaare (2 Meter lang)

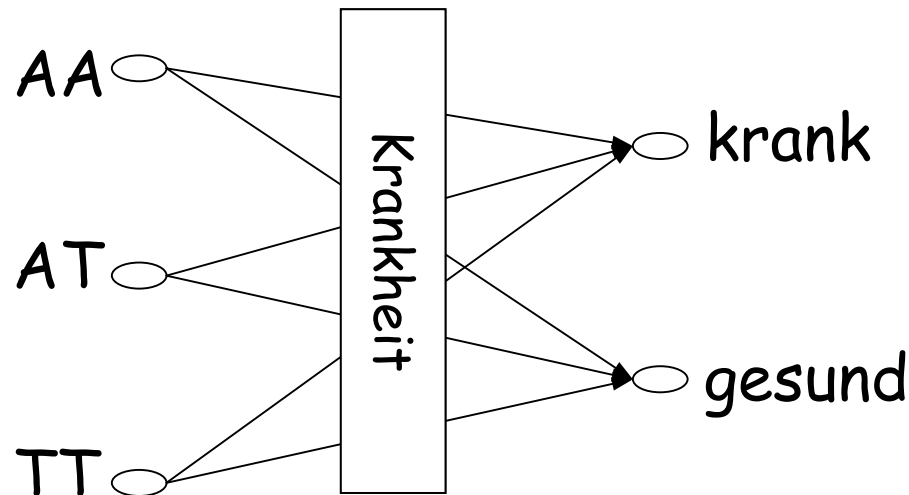
...GCTAAATGCGCCGTACT...



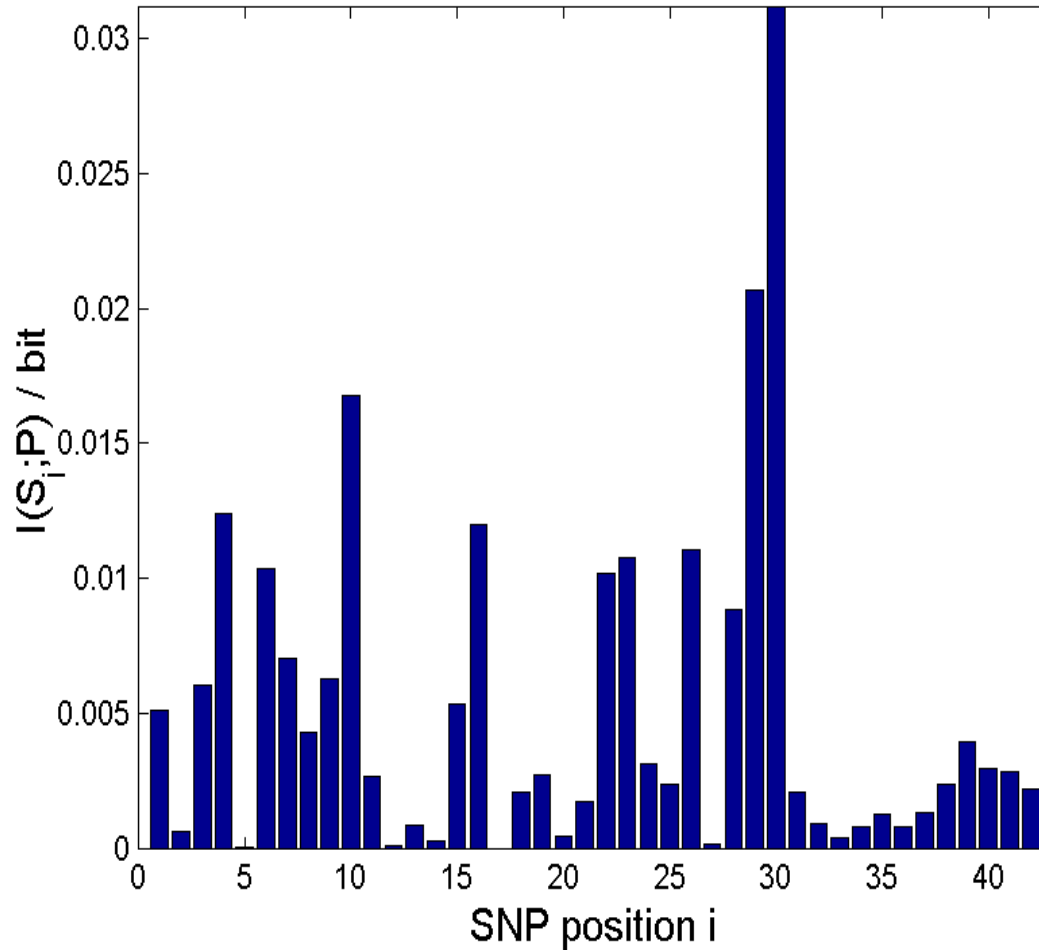
Anwendung der Informationstheorie bei der Suche nach Krankheit verursachenden Stellen im Genom

- Single nucleotide polymorphisms (SNPs)
 - Stellen im Genom mit größerer Variation als 1%

Position des SNPs	1	2	3	4	5	6	7	8	9
Genom der Mutter	A	A	T	A	T	T	T	T	T
Genom des Vaters	A	T	T	A	A	T	A	A	T

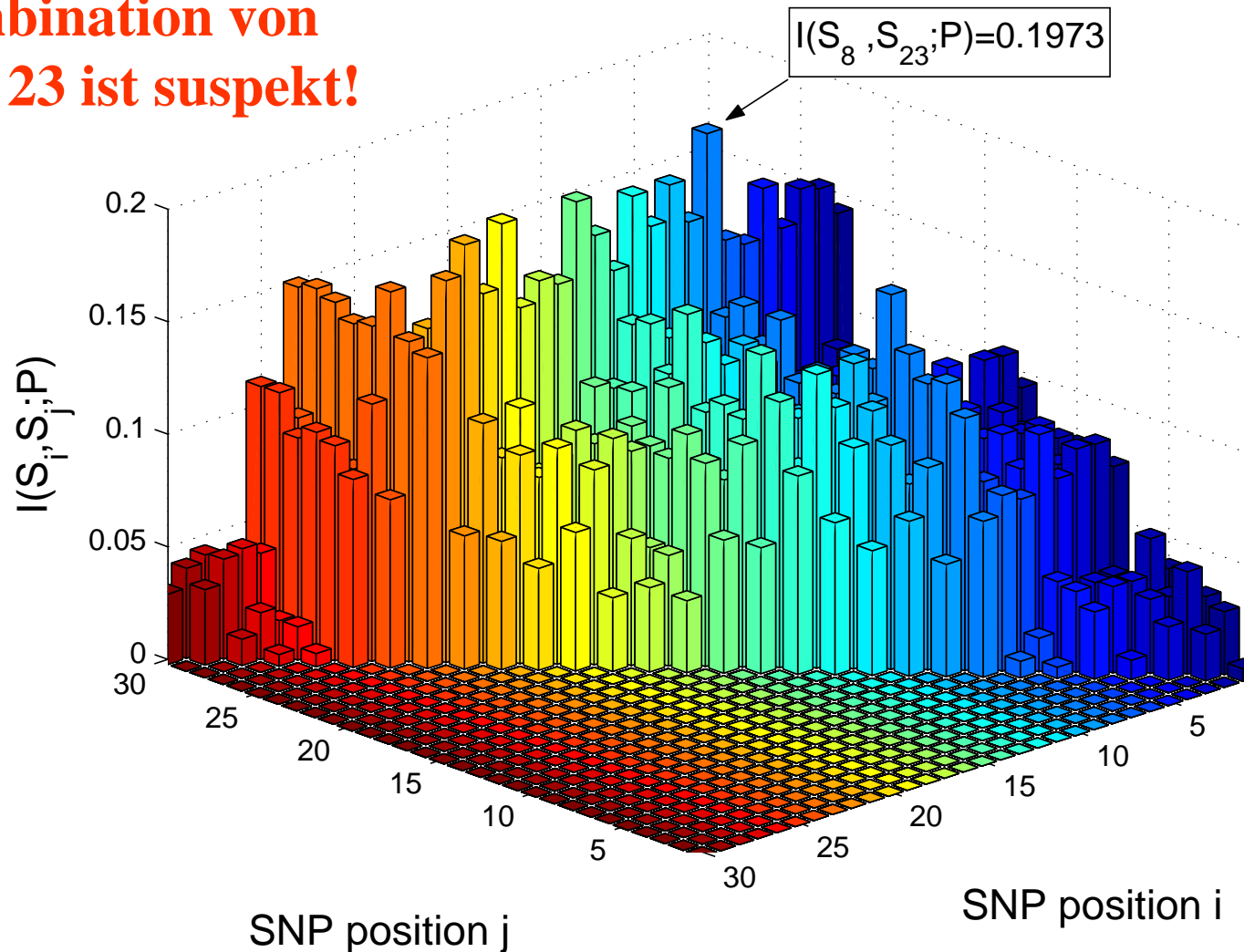


Schizophrenie: verdächtiger SNP

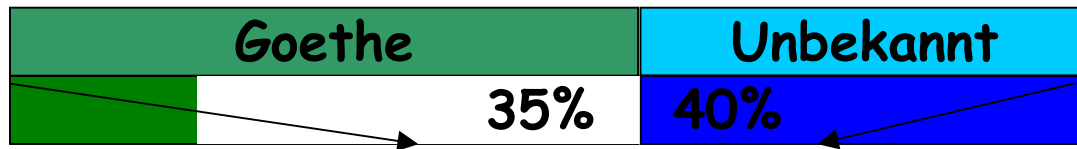
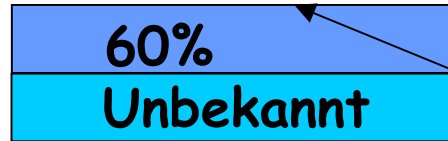


Informationsübertragung von 2 SNPs zu Parkinson

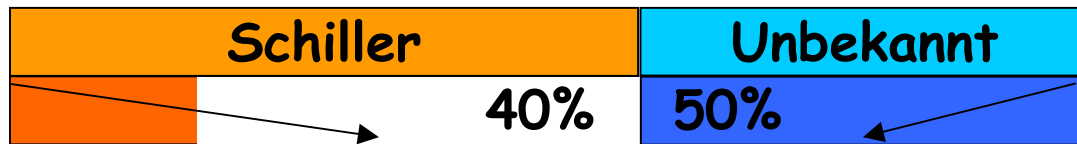
**Eine Kombination von
SNP 8 und 23 ist suspekt!**



Distanzberechnung von Texten bzw. DNS Sequenzen mittels Kompression



Distanz zu Goethe = $\frac{40\%}{60\%}$

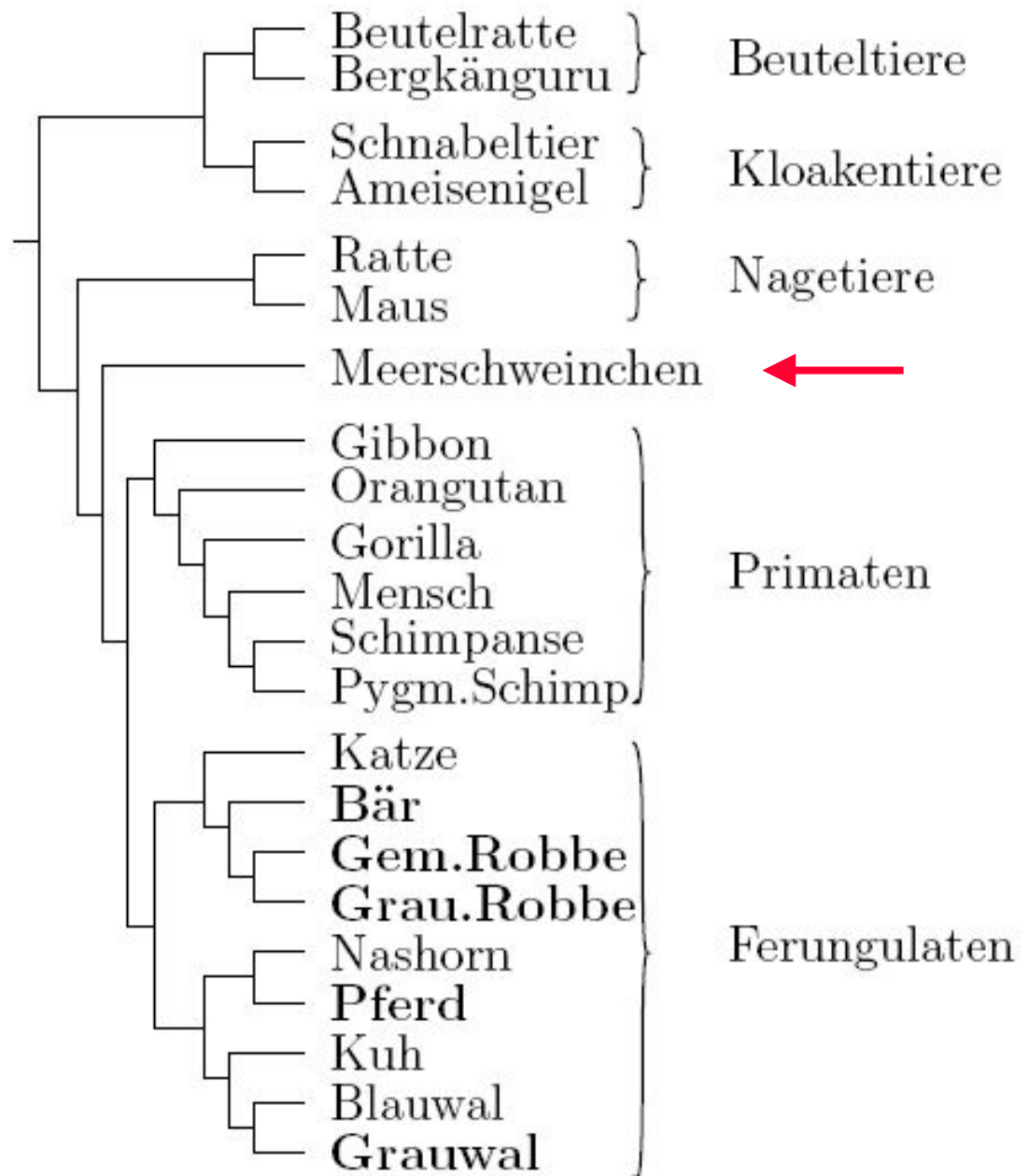


Distanz zu Schiller = $\frac{50\%}{60\%}$

Kompressionsdistanzen zwischen verschiedenen Tierarten mtDNS etwa 16.000 Basen

Distanz	Wal	Gemeine Robbe	Graue Robbe	Pferd	Bär
Wal	0	0,85	0,84	0,81	0,88
Gemeine Robbe	0,85	0	0,22	0,77	0,74
Graue Robbe	0,84	0,22	0	0,78	0,74
Pferd	0,81	0,77	0,78	0	0,86
Bär	0,88	0,74	0,74	0,86	0

Baum der Tierarten abgeleitet aus der Shannonschen Kompressionstheorie



Folgerungen:

- Baum stimmt mit anderen Forschungen überein (Wallace)
- Meerschweinchen ist umstritten

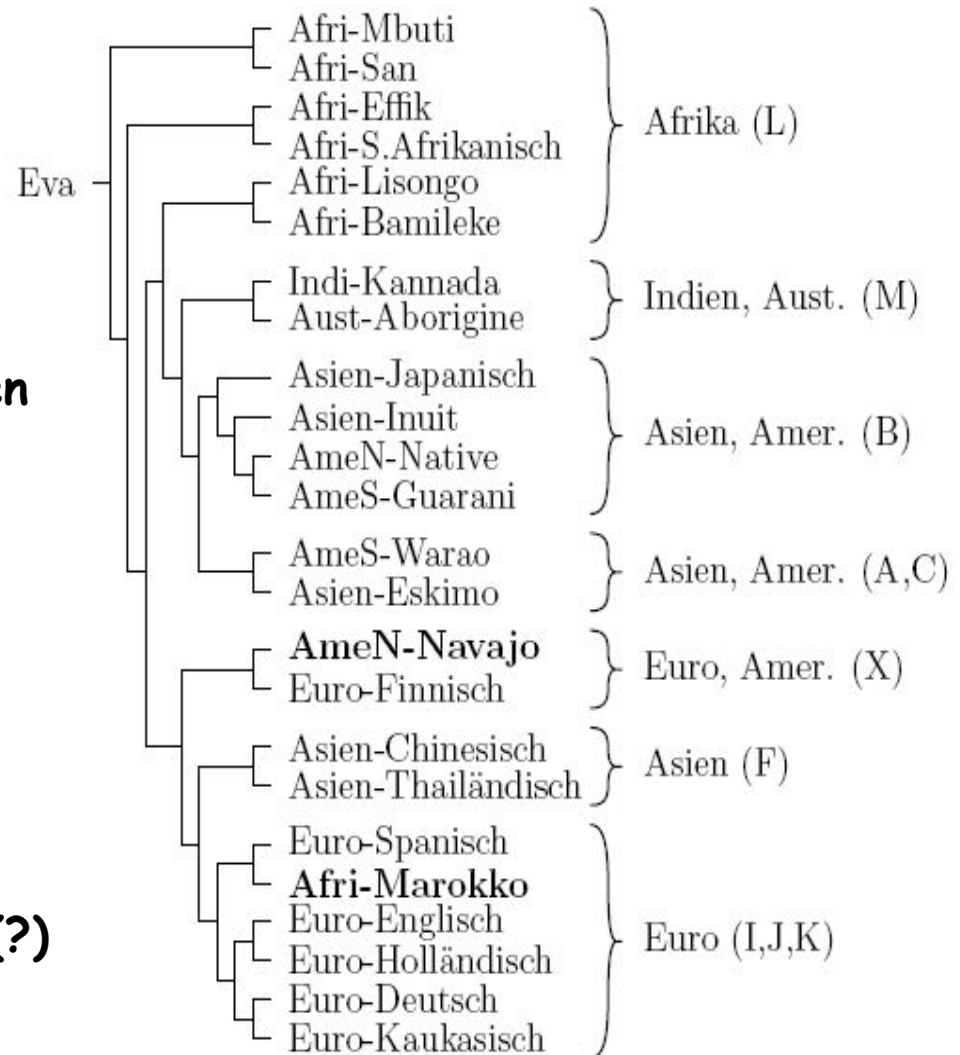
Baum des homo sapiens abgeleitet aus der Shannonschen Kompressionstheorie

Konstruiert aus der mtDNS

Kompressor = DNACompress

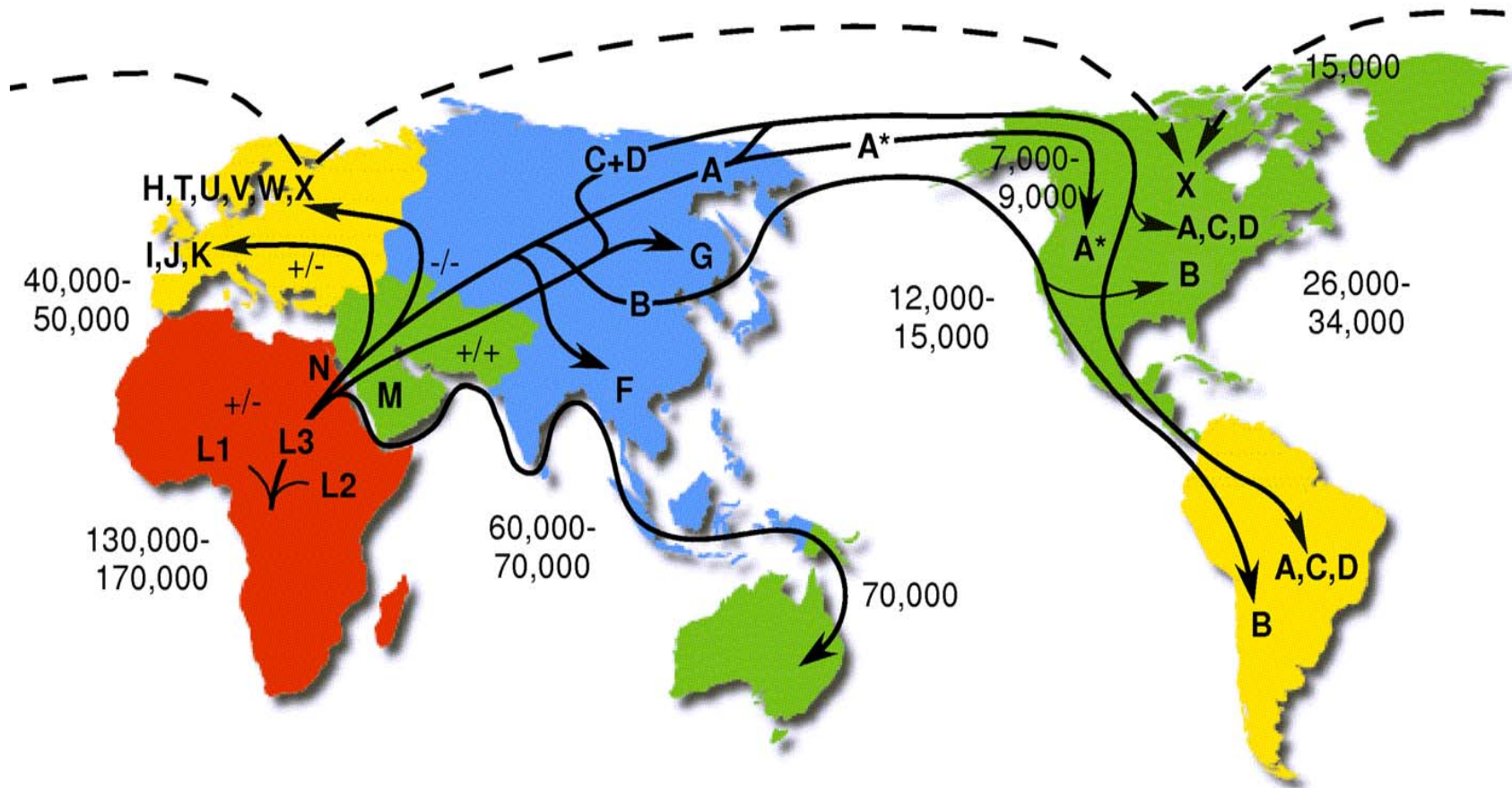
Folgerungen:

- Baum stimmt mit anderen Forschungen überein (Cavalli et al)
- Ursprung der Menschheit liegt in Afrika
- Marokko wurde über Europa besiedelt
- Besiedelung Nordamerikas erfolgte über Alaska und Finnland(?)



Migrationsmuster der Menschheit

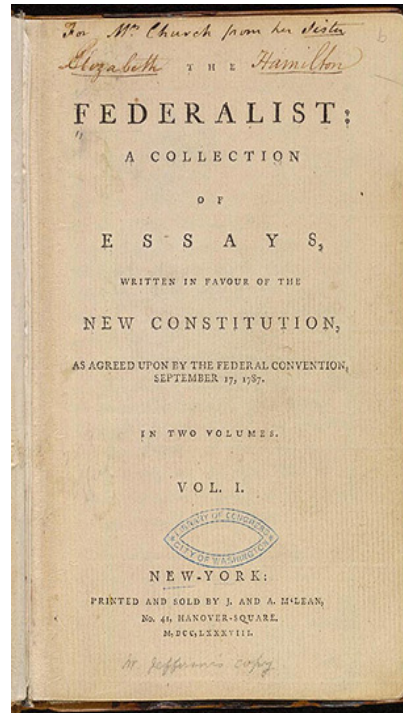
Copyright 2002 © Mitomap.org



Autorenschaft Erkennung

Federalist Papers

- 1787-1788 (New York)
- 85 Aufsätze
(Pseudonym PUBLIUS)
- geschrieben von:
 - J. Madison
 - A. Hamilton
 - J. Jay
- bei 12 Aufsätzen ist die Autorenschaft strittig



49	✓	7,9%	17%
50	✓	4,8%	16%
51	✓	7,3%	15%
52	✓	6,4%	17%
53	✓	5,8%	13%
54	✓	3,6%	13%
55	✓	4,8%	15%
56	✓	5,4%	14%
57	✓	1,0%	14%
58	✓	4,2%	16%
62	✓	5,3%	12%
63	✓	4,7%	13%

Informationstheorie ist eine mathematisch anspruchsvolle Theorie.

Eine Seite aus den „Transactions on Information Theory“:

1126

IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 47, NO. 3, MARCH 2001

In (95), the second term will be

$$\begin{aligned}
 & \sum_{m_0=1}^N \sum_{j=1}^{m_0} \int_{\mathbf{n} \in L_j(m_0, s^*)} \left(\sum_{i=1}^{m_0} v_i u_i - v_j \right) dF(\mathbf{n}) \\
 & > \sum_{m_0=1}^N \sum_{j=1}^{m_0} \int_{\mathbf{n} \in L_j(m_0, s^*)} \\
 & \quad \left(\sum_{i \neq j, i=1}^{m_0} [v_j + (\lambda_i - \lambda_j) s^*] u_i + v_j u_j - v_j \right) dF(\mathbf{n}) \\
 & = \sum_{m_0=1}^N \sum_{j=1}^{m_0} \int_{\mathbf{n} \in L_j(m_0, s^*)} \left(\sum_{i \neq j, i=1}^{m_0} (\lambda_i - \lambda_j) s^* u_i \right) dF(\mathbf{n}) \\
 & = \sum_{m_0=1}^N \sum_{j=1}^{m_0} \int_{\mathbf{n} \in L_j(m_0, s^*)} \\
 & \quad \left(\sum_{i \neq j, i=1}^{m_0} \lambda_i s^* u_i - \lambda_j s^* [1 - u_j] \right) dF(\mathbf{n}) \\
 & = s^* \sum_{m_0=1}^N \sum_{j=1}^{m_0} \int_{\mathbf{n} \in L_j(m_0, s^*)} \\
 & \quad \left(\sum_{i \neq j, i=1}^m \lambda_i u_i - \lambda_j [1 - u_j] \right) dF(\mathbf{n}). \tag{98}
 \end{aligned}$$

c) $\forall \mathbf{n} \in \tilde{L}_j(m_0, s^*), j = 1, 2, \dots, m_0$, since $\frac{1}{s^*} = z_j = x_{j-1}$, we have

$$v_{j-1} = v_j + (\lambda_{j-1} - \lambda_j) s^*.$$

$$\begin{aligned}
 & = \sum_{m_0=1}^N \sum_{j=1}^{m_0} \left[\left(\sum_{i \neq j, i=1}^{m_0} \lambda_i s^* u_i \right) - \lambda_j s^* (1 - u_j) \right. \\
 & \quad \left. + (\lambda_j - \lambda_{j-1}) (1 - \tau^*) s^* \right] \Pr(\tilde{L}_j(m_0, s^*)) \\
 & = \sum_{m_0=1}^N \sum_{j=1}^{m_0} \left\{ \left(\sum_{i \neq j-1, i \neq j, i=1}^{m_0} \lambda_i s^* u_i \right) + \lambda_j s^* (u_j - \tau^*) \right. \\
 & \quad \left. + \lambda_{j-1} s^* [u_{j-1} - (1 - \tau^*)] \right\} \\
 & \quad \times \Pr(\tilde{L}_j(m_0, s^*)) \\
 & = s^* \sum_{m_0=1}^N \sum_{j=1}^{m_0} \left\{ \left(\sum_{i \neq j-1, i \neq j, i=1}^{m_0} \lambda_i u_i \right) + \lambda_j (u_j - \tau^*) \right. \\
 & \quad \left. + \lambda_{j-1} [u_{j-1} - (1 - \tau^*)] \right\} \\
 & \quad \times \Pr(\tilde{L}_j(m_0, s^*)). \tag{99}
 \end{aligned}$$

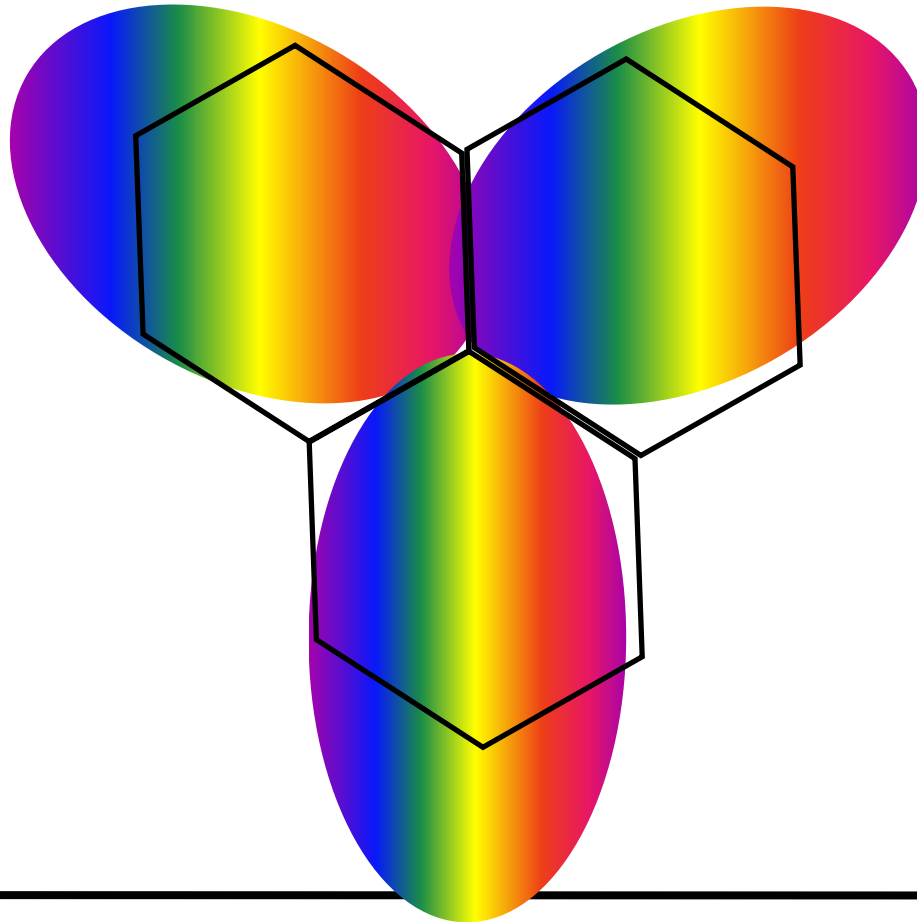
Therefore, substituting (97)–(99) into (95), we have

$$\begin{aligned}
 & \sum_{m_0=1}^N E_{\mathbf{n} \in \Omega_{m_0}} \left[\sum_{i=1}^{m_0} v_i u_i \right] - \bar{P} \\
 & > s^* \sum_{m_0=1}^N \left\{ \int_{\mathbf{n} \in A(m_0)} \sum_{i=1}^{m_0} \lambda_i u_i dF(\mathbf{n}) \right. \\
 & \quad \left. - \int_{\mathbf{n} \in B(m_0)} \sum_{i=1}^{m_0} \lambda_i u_i dF(\mathbf{n}) \right\}
 \end{aligned}$$

Die Grenzen der Shannonschen Informationstheorie:

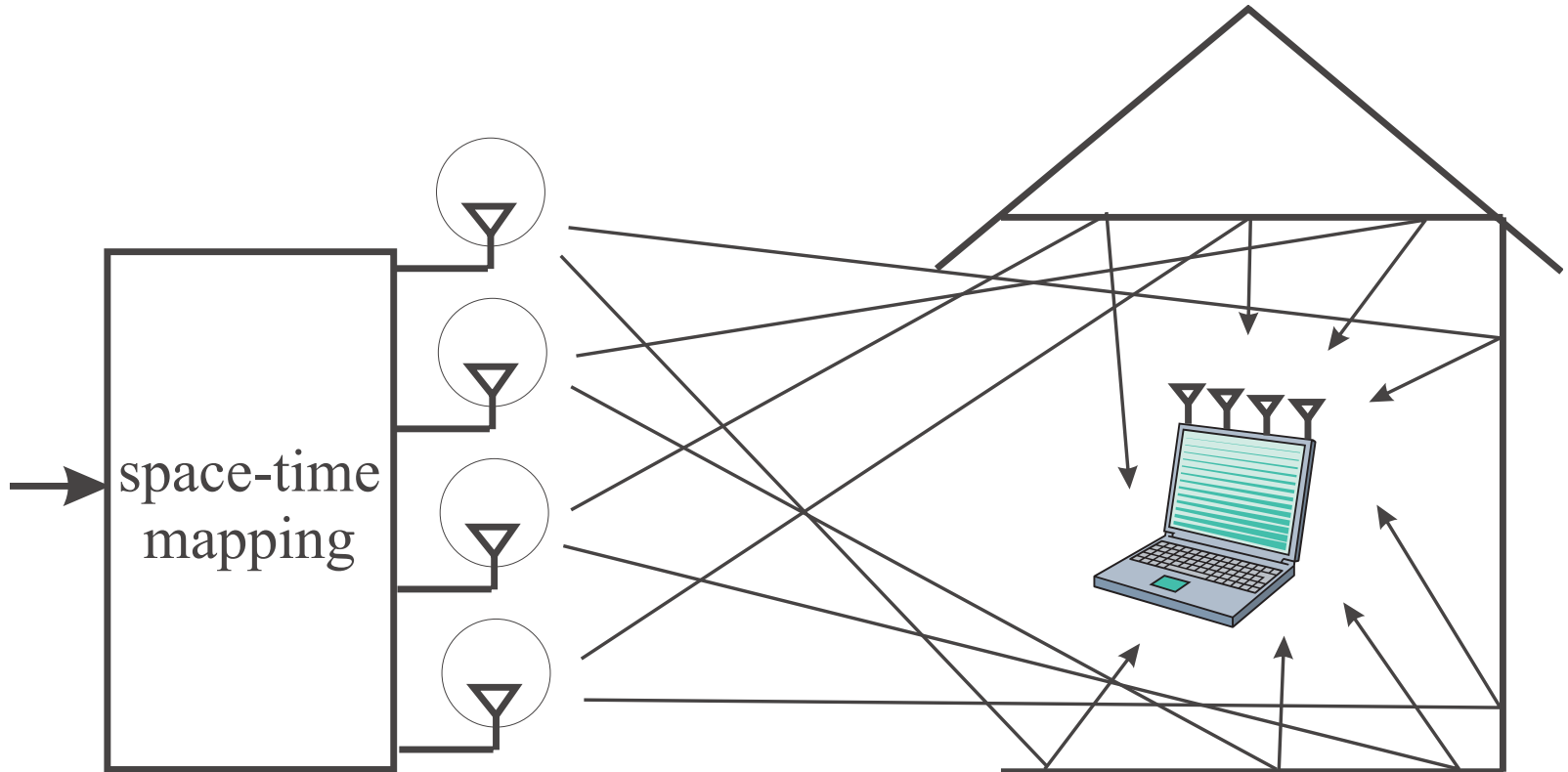
Offene Fragen sind z.B.:

Welchen Datenverkehr kann man in einer Mobilfunkzelle unterbringen?

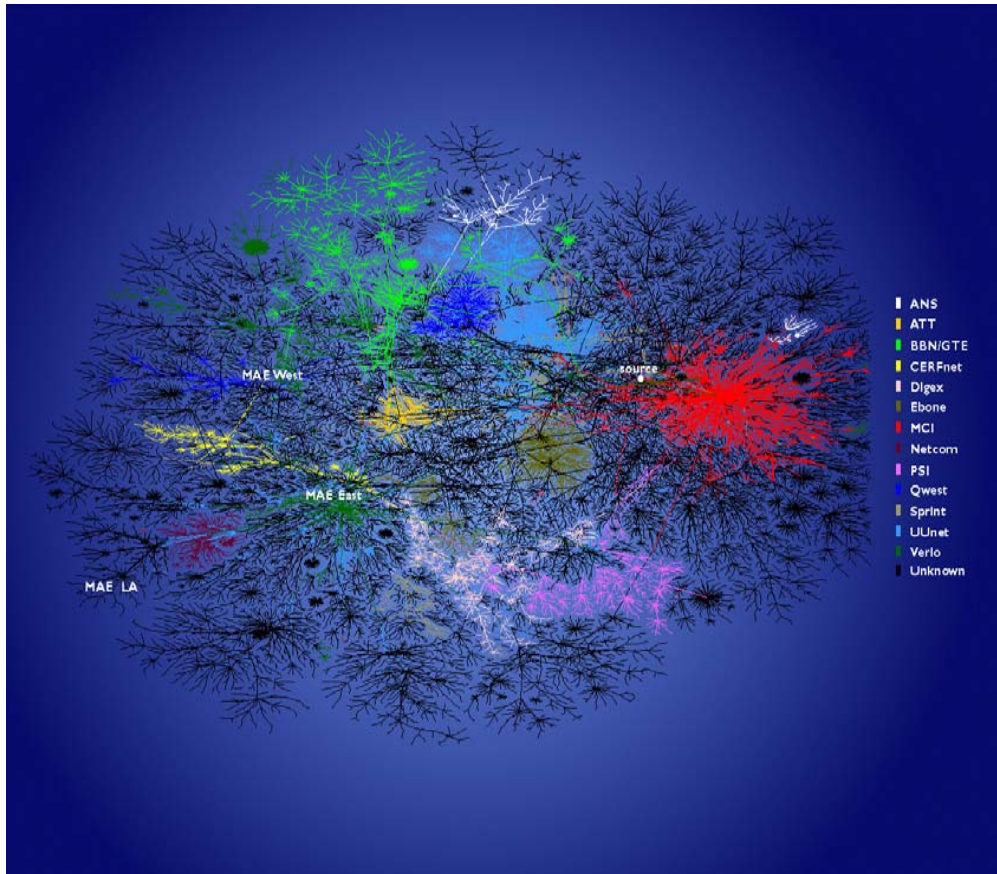


Offene Fragen sind z.B.:

Informationstheorie bei Mehrantennen-Systemen



Die Grenzen der Shannonschen Informationstheorie: Offene Fragen sind z.B.: Was ist die theoretische Kapazität des Internets?



Aktuelles Forschungsgebiet: Netzwerkinformationstheorie

Die Grenzen der Shannonschen Informationstheorie:

Shannon berücksichtigt nicht den Inhalt
(die Semantik) der Information.

Er sagt 1948:

„The fundamental problem of communication is that of reproducing at one point a message selected at another point. Frequently the messages have meaning... These semantic aspects of communication are irrelevant to the engineering problem“

Die Grenzen der Shannonschen Informationstheorie:

Weisheit

Wissen

Information

Daten

T.S Eliot, „The Rock“:

Where is the Life we have lost in living?

Where is the wisdom we have lost in knowledge?

Where is the knowledge we have lost in information?

