「情報とは何か」を考えてみる

モデリング研究系　伊庭幸人

「情報」の学問とは？

うちの父親（大正生まれ）
　ん～「情報」ってやつはわからん

いろんな答えがありうる

いろいろ疑問

統計学は「情報」に関係あるの？

統数研の人はなぜここに集まっているの？

シャノンの情報理論というのが
　あるそうだが，それって
　　「情報」に関係があるの？

ここでの主張：「情報」の科学とは

知識の表現や抽出・推論の方法を考え，
予測や発見に役立てる学問

広い意味での統計科学
人工知能　パターン認識　機械学習
疫学　データマイニング　情報理論
統計学　・・・　みんな本質を共有

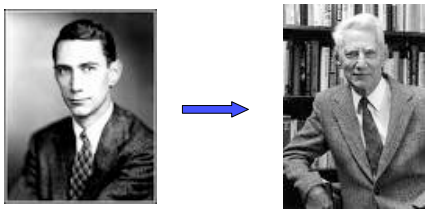
情報圧縮

シャノンの情報理論2本の柱のひとつ

@可逆圧縮(lha,zip, gif,png, mpegの一部)
　テキストの圧縮　画像も一部はこれ

@非可逆圧縮（jpg, mpegの大半）
　画像では主流

シャノン



この写真は2枚とも非可逆圧縮

夢の可逆圧縮は可能か？

すべてのファイルを1/100のサイズに圧縮します

詐欺

長さ1000ビットのファイル　2^{1000} 個
長さ10　ビットのファイル　2^{10} 個
長さ999　ビットのファイル　2^{999} 個

N個のものをN-1個に入れたら？

必ずどれか重複する
⇒ 可逆圧縮ではありえない

「鳥の巣箱論法」
椅子のほうが人より少なければ
誰か座れない人が出る

かならず人のほうががあまる

なぜ可逆圧縮できるか

原理
- 出現確率の低い対象には長いコードを
- 出現確率の高い対象には短いコードを

割り当てればよい

全部短くする　→　区別ができなくなるからだめ

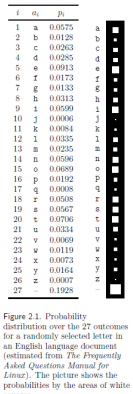
古典的な例：　英字の頻度

e	11.4	12.3
t	8.2	9.1
a	8.4	8.1

多いほう(%)

j	0.21	0.2
q	0.08	0.1
z	0.08	0.07

少ないほう(%)



i	a_i	p_i
1	a	0.0575
2	b	0.0128
3	c	0.0263
4	d	0.0285
5	e	0.0913
6	f	0.0173
7	g	0.0133
8	h	0.0313
9	i	0.0599
10	j	0.0006
11	k	0.0084
12	l	0.0335
13	m	0.0235
14	n	0.0596
15	o	0.0689
16	p	0.0192
17	q	0.0008
18	r	0.0508
19	s	0.0567
20	t	0.0706
21	u	0.0334
22	v	0.0069
23	w	0.0119
24	x	0.0073
25	y	0.0164
26	z	0.0007
27		0.1928

Figure 2.1. Probability distribution over the 27 outcomes for a randomly selected letter in an English language document (estimated from *The Frequently Asked Questions Manual for Linux*). The picture shows the probabilities by the areas of white squares.

文字の出やすさ

David MacKay
Information theory,
Inference and Learning
Algorithm より

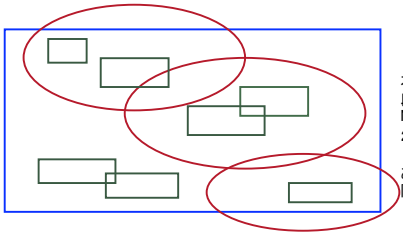
モールス符号

欧　　文

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

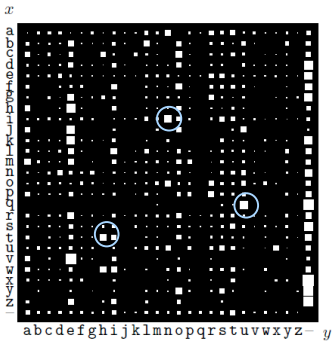
文字の相関

- THE とかHEとかいう言葉がたくさんある
- ⇒　Hのあとは**E**が多いはず



本，記事や章，
段落，文，単語，
N文字，. . . , 3文字，
2文字

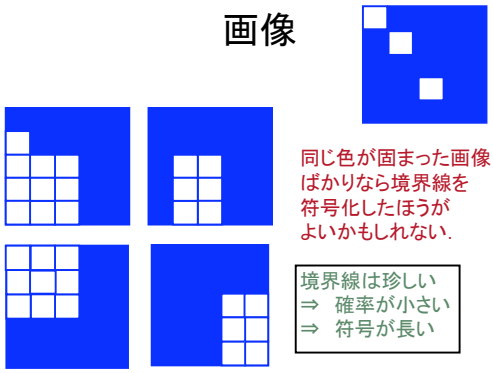
あらゆるレベルで
階層的に相関構造



(a) $P(y|x)$

2文字の相関

画像



同じ色が固まった画像
ばかりなら境界線を
符号化したほうが
よいかもしれない．

境界線は珍しい
⇒　確率が小さい
⇒　符号が長い

統計科学の仕事

世の中のこういう「構造」をどう表現したらよいか

確率モデルによる世の中の構造の表現と発見

情報圧縮に限らずあらゆる予測・分析・制御のために

そこで重要なポイントは？

細かくみすぎるとだめ！

「細かくみすぎると，だめになる」
ことは世の中にたくさんある

非独立性の表現

相関構造の表現の一番かんたんな
方法は「ブロック」の頻度を考えること

b a a a b a a a a a b a b a a b b a b a
ba aa ba aa aa ba ba ab ba ba
baa aba aaa aba baa bba ba

どんどんブロックを大きくすると・・

0 1 0 0 1 0 0 1 0 1 0 1 0 0 0 0
01 00 10 01 01 01 00 00
010 010 010 101 000 0
0100 1001 0101 0000
01001 00101 01000 0
・・・・
0100100101010000

ん？

辞書を忘れてはいけない

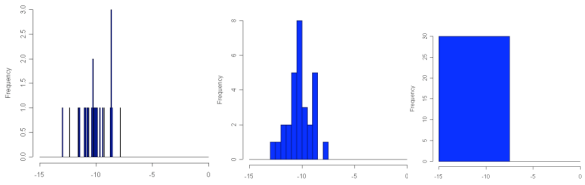
解読するためには辞書が必要

- ブロック長　＝　データ全体の長さ
- 辞書　＝　もとのデータをそのまま含む

⇒　あきらかに無意味（すでに知ってる！）
⇒　データ数が有限なら複雑にしすぎてもだめ

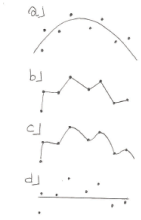
ヒストグラムの切り方

× ○ ×



過剰学習（overfit）

曲線をあてはめる



社会調査の解析

100項目　（YES, NO）

2の100乗とおり

100次元の表のなかみは0か
1ばかりになりかねない！

どの人もこの世で唯一の人

汎化（generalization）

ロボットを進化させる

例：歩き方の学習
　プログラムをランダムに変更　⇒　よいものを選ぶ

何が学習されているのか？

@　歩き方一般を学習
@　訓練に使った廊下の構造を学習
@　その廊下にあるごみの配置を学習

学習能力が高いほど細部まで丸暗記してしまい
かえって悪くなる可能性がある

「単純さ」の論理

なぜ単純なモデルが好まれる？
なぜ「規則」と「雑音」「偶然」に分ける？

- MDL　情報圧縮の上でそれが有利だから
- AIC　予測のためにそれが有利だから

そのほかにも，いろいろな考え方がある

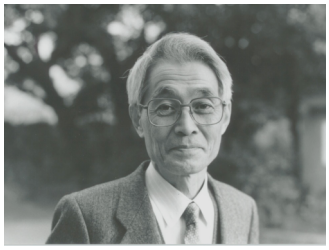
統計科学の挑戦

対象の個別性・多様性をできるだけ捨てずに
意味のある規則性を抽出し，
予測・情報圧縮などを行う

知識の表現の工夫（モデリング）＋単純さの評価

たとえば
@ 平均的な客に合わせた販売 ⇒ 多様な客層，曜日や季節の違い‥
@ どの人も同じ薬 ⇒ ひとりひとりの個性を考えた投薬
@ 「生えている場所」を無視した植物の研究 ⇒ 空間変数のとりこみ
@ いつでも同じ通信のやり方 ⇒ 場所の移動やその日の状態によって‥

赤池先生（AICの生みの親）



情報のありか

「情報」はいろんなところに隠れている

情報じゃない，と思ったところが情報だったりする

文章の著者の判定

金明哲氏の統計数理研究所での学位論文

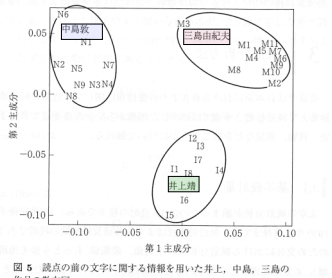


図 5　読点の前の文字に関する情報を用いた井上，中島，三島の作品の散布図

私は，そうは思わない

しかし，それは違う

〇〇であるが，そう思う

読点の直前の文字の頻度を調べる

ディジタル透かし

0～255の濃度
の画像があったとする

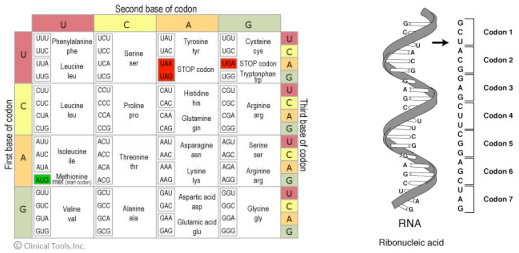
122 122 123 123 124 125
122 122 123 123 124 125
122 122 123 123 124 125

123 123 122 123 124 125
122 122 123 123 124 125
122 122 123 123 124 125

122 122 123 123 124 125
122 122 123 123 124 125
122 122 123 123 124 125

123 123 122 123 124 125
122 122 123 123 124 125
122 122 123 123 124 125

遺伝暗号：同義コドン



アミノ酸との対応では3つめは通常は意味がない

数理的に考えること

数式をいじったり，数字を計算するだけが
数理じゃない！

隠された要素を見抜くことの重要性