Science is to test crazy ideas – Engineering is to put these ideas into Business

Andreas Holzinger
VO 709.049 Medical Informatics
19.10.2016 11:15-12:45

Lecture 02 Back to the Future – Fundamentals of biomedical Data, Information, and Knowledge

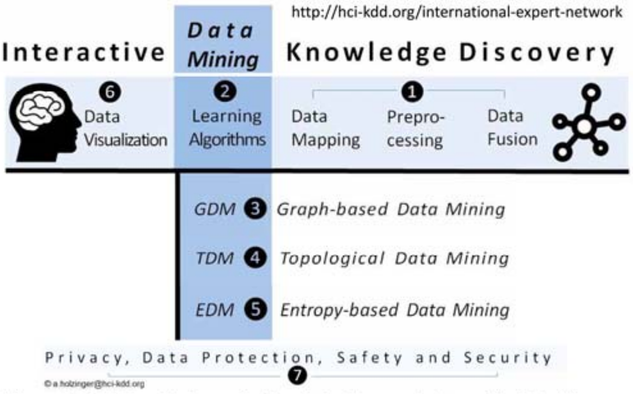
a.holzinger@tugraz.at
Tutor: markus.plass@student.tugraz.at
http://hci-kdd.org/biomedical-informatics-big-data

Schedule

- 1. Introduction: Computer Science meets Life Sciences, challenges and future directions
- **2. Back to the future: Fundamentals of Data, Information and Knowledge**
- 3. Structured Data: Coding, Classification (ICD, SNOMED, MeSH, UMLS)
- 4. Biomedical Databases: Acquisition, Storage, Information Retrieval and Use
- 5. Semi structured and weakly structured data (structural homologies)
- 6. Multimedia Data Mining and Knowledge Discovery
- 7. Knowledge and Decision: Cognitive Science & Human-Computer Interaction
- 8. Biomedical Decision Making: Reasoning and Decision Support
- 9. Intelligent Information Visualization and Visual Analytics
- 10. Biomedical Information Systems and Medical Knowledge Management
- 11. Biomedical Data: Privacy, Safety and Security
- 12. Methodology for Information Systems: System Design, Usability and Evaluation

ML needs a concerted effort fostering integrated research

http://hci-kdd.org/international-expert-network



Interactive Data Mining Knowledge Discovery

6 Data Visualization

2 Learning Algorithms

1 Data Mapping, Prepro-cessing, Data Fusion

GDM 3 Graph-based Data Mining

TDM 4 Topological Data Mining

EDM 5 Entropy-based Data Mining

Privacy, Data Protection, Safety and Security 7

© a.holzinger@hci-kdd.org

Holzinger, A. 2014. Trends in Interactive Knowledge Discovery for Personalized Medicine: **Cognitive Science meets Machine Learning.** IEEE Intelligent Informatics Bulletin, 15, (1), 6-14.

Keywords

- Data
- Information
- Knowledge
- Dimensionality of data
- Information complexity
- Information (Shannon) entropy
- Mutual Information/Cross Entropy
- Kullback-Leibler Divergence

Learning Goals

- … be aware of the types and categories of different data sets in biomedical informatics;
- … know some differences between data, information, and knowledge;
- … be aware of standardized/non-standardized and well-structured/"un-structured" information/data;
- … have a basic overview on information theory and the concept of information entropy;
- … are aware of the importance of the Kullback-Leibler divergence

Advance Organizer (1/2)

- **Abduction** = cyclical process of generating possible explanations (i.e., identification of a set of hypotheses that are able to account for the clinical case on the basis of the available data) and testing those (i.e., evaluation of each generated hypothesis on the basis of its expected consequences) for the abnormal state of the patient at hand;
- **Abstraction** = data are filtered according to their relevance for the problem solution and chunked in schemas representing an abstract description of the problem (e.g., abstracting that an adult male with haemoglobin concentration less than 14g/dL is an anaemic patient);
- **Artefact/surrogate** = error or anomaly in the perception or representation of information trough the involved method, equipment or process;
- **Data** = physical entities at the lowest abstraction level which are, e.g. generated by a patient (patient data) or a (biological) process; data contain no meaning;
- **Data quality** = Includes quality parameter such as : Accuracy, Completeness, Update status, Relevance, Consistency, Reliability, Accessibility;
- **Data structure** = way of storing and organizing data to use it efficiently;
- **Deduction** = deriving a particular valid conclusion from a set of general premises;
- **DIK-Model** = Data-Information-Knowledge three level model
- **Disparity** = containing different types of information in different dimensions
- **Heart rate variability (HRV) =** measured by the variation in the beat-to-beat interval;
- **HRV artifact** = noise through errors in the location of the instantaneous heart beat, resulting in errors in the calculation of the HRV, which is highly sensitive to artifact and errors in as low as 2% of the data will result in unwanted biases in HRV calculations;

Advance Organizer (2/2)

- **Induction** = deriving a likely general conclusion from a set of particular statements;
- **Information** = derived from the data by interpretation (with feedback to the clinician);
- **Information Entropy** = a measure for uncertainty: highly structured data contain low entropy, if everything is in order there is no uncertainty, no surprise, ideally H = 0
- **Knowledge** = obtained by inductive reasoning with previously interpreted data, collected from many similar patients or processes, which is added to the "body of knowledge" (explicit knowledge). This knowledge is used for the interpretation of other data and to gain implicit knowledge which guides the clinician in taking further action;
- **Large Data** = consist of at least hundreds of thousands of data points
- **Multi-Dimensionality** = containing more than three dimensions and data are multivariate
- **Multi-Modality** = a combination of data from different sources
- **Multivariate** = encompassing the simultaneous observation and analysis of more than one statistical variable;
- **Reasoning** = process by which clinicians reach a conclusion after thinking on all facts;
- **Spatiality** = contains at least one (non-scalar) spatial component and non-spatial data
- **Structural Complexity** = ranging from low-structured (simple data structure, but many instances, e.g., flow data, volume data) to high-structured data (complex data structure, but only a few instances, e.g., business data)
- **Time-Dependency** = data is given at several points in time (time series data)
- **Voxel** = volumetric pixel = volumetric picture element

Glossary (incomplete)

- ApEn = Approximate Entropy;
- DIK = Data-Information-Knowledge-3-Level Model;
- GraphEn = Graph Entropy;
- H = Entropy (General);
- HRV = Heart Rate Variability;
- MaxEn = Maximum Entropy;
- MinEn = Minimum Entropy;
- NE = Normalized entropy (measures the relative informational content of both the signal and noise);
- PDB = Protein Data Base;
- SampEn = Sample Entropy;

Common Mathematical Notations

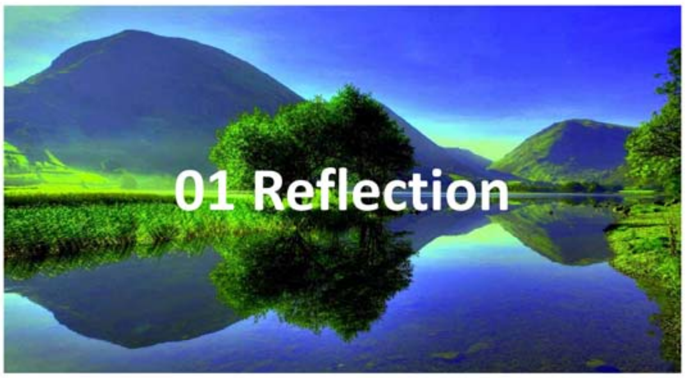
"In mathematics you don't understand things. You just get used to them" – John von Neumann

Data	
n	Number of samples
d	Number of input variables
$\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]$	Matrix of input samples
$\mathbf{y} = [y_1, \dots, y_n]$	Vector of output samples
$\mathbf{Z} = [\mathbf{X}, \mathbf{y}]$	Combined input–output training data or
$\mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_n]$	Representation of data points in a feature space
Distribution	
P	Probability
$F(\mathbf{x})$	Cumulative probability distribution function (cdf)
$p(\mathbf{x})$	Probability density function (pdf)
$p(\mathbf{x}, y)$	Joint probability density function
$p(\mathbf{x}; \omega)$	Probability density function, which is parameterized
$p(y\|\mathbf{x})$	Conditional density
$t(\mathbf{x})$	Target function

Mathematical Notations for Course LV 185.A83 Machine Learning for Health Informatics

Agenda for today

- **01 Reflection – follow-up from last lecture**
- **02 What is data?**
- **03 Excursus: Data Integration – Data Fusion**
- **04 What is Information?**
- **05 What is Knowledge?**
- **06 A clinical view on data, information, and knowledge**



01 Reflection

Image source: http://www.hutui6.com/reflection-wallpapers.html

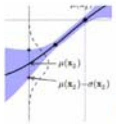
Warm-up Quiz

1

2

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta) * p(\theta)}{p(\mathcal{D})}$$

3



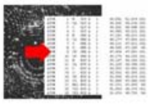
4

5

6

4. Check
3. iML
2. Preprocessing
1. Input

7



8

9

Solutions in the Appendix

Question: Where is the Biologist in this image?

Image source: http://www.efmc.info/medchemwatch-2014-1/lab.php

Domingos, P. 2015. The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World, Penguin UK.

Repetition of Bayes - on the work of Laplace

What is the simplest mathematical operation for us?

$$p(x) = \sum_x (p(x,y)) \qquad (1)$$

How do we call repeated adding?

$$p(x,y) = p(y|x) * p(y) \qquad (2)$$

Laplace (1773) showed that we can write:

$$p(x,y) * p(y) = p(y|x) * p(x) \qquad (3)$$

Now we introduce a third, more complicated operation:

$$\frac{p(x,y) * p(y)}{p(y)} = \frac{p(y|x) * p(x)}{p(y)} \qquad (4)$$

We can reduce this fraction by $p(y)$ and we receive what is called Bayes rule:

$$p(x,y) = \frac{p(y|x) * p(x)}{p(y)} \qquad p(h|d) = \frac{p(d|h)p(h)}{p(d)} \qquad (5)$$

Problem: Diagnoses

- Your MD has bad news and good news for you.
- Bad news first: You are tested positive for a serious disease, and the test is 99% accurate (T)
- Good news: It is a rare disease, striking 1 in 10,000 (D)
- **How worried would you now be?**

$$posterior\ p(x) = \frac{likelyhood * prior\ p(x)}{evidence} \qquad p(h|d) = \frac{p(d|h)p(h)}{p(d)}$$

$$p(T=1|D=1) = p(d|h) = 0{,}99 \ and$$
$$p(D=1) = p(h) = 0{,}0001$$

$$p(D=1 \,\big|\, T=1) = \frac{(0{,}99)*(0{,}0001)}{(1-0{,}99)*(1-0{,}0001)+0{,}99*0{,}0001} =$$

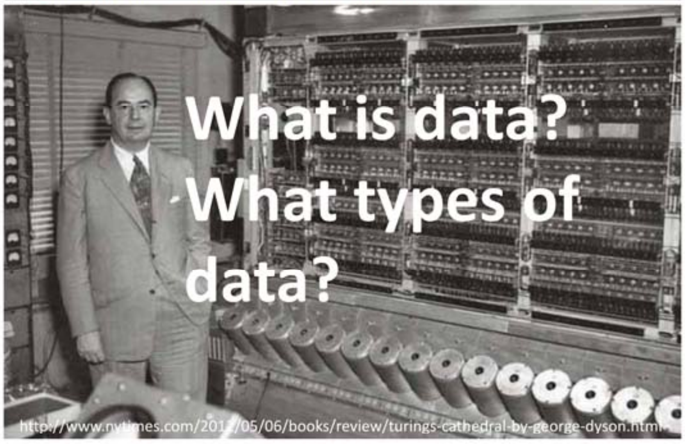
$$\mathbf{= 0{,}0098}$$

Key Problems

- Heterogeneous, distributed, inconsistent data sources (need for **data integration** & fusion) [1]
- **Complex data** (high-dimensionality – challenge of dimensionality reduction and visualization) [2]
- Noisy, uncertain, missing, dirty, and imprecise, imbalanced data (challenge of **pre-processing**)
- The discrepancy between data-information-knowledge (**various definitions**)
- **Big data** sets (manual handling of the data is awkward, and often impossible) [3]

1. Holzinger A, Dehmer M, & Jurisica I (2014) Knowledge Discovery and interactive Data Mining in Bioinformatics - State-of-the-Art, future challenges and research directions. BMC Bioinformatics 15(S6):I1.
2. Hund, M., Sturm, W., Schreck, T., Ullrich, T., Keim, D., Majnaric, L. & Holzinger, A. 2015. Analysis of Patient Groups and Immunization Results Based on Subspace Clustering. In: LNAI 9250, 358-368.
3. Holzinger, A., Stocker, C. & Dehmer, M. 2014. Big Complex Biomedical Data: Towards a Taxonomy of Data. in CCIS 455. Springer 3-18.

Institute for Advanced Study, Princeton University



What is data? What types of data?

http://www.nytimes.com/2012/05/06/books/review/turings-cathedral-by-george-dyson.html

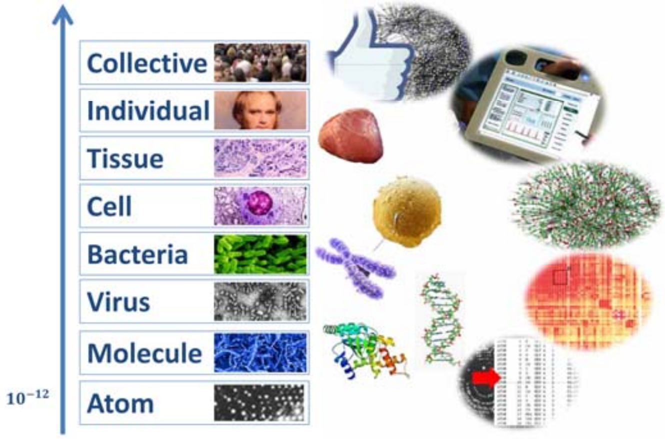
Traditional Statistics versus Machine Learning

- Data in traditional Statistics
- Low-dimensional data ($< \mathbb{R}^{100}$)
- Problem: Much noise in the data
- Not much structure in the data but it can be represented by a simple model

- Data in Machine Learning
- High-dimensional data ($\gg \mathbb{R}^{100}$)
- Problem: not noise , but complexity
- Much structure, but the structure but can **not** be represented by a simple model

Lecun, Y., Bengio, Y. & Hinton, G. 2015. Deep learning. Nature, 521, (7553), 436-444.

Slide 2-1: Biomedical Data Sources



Data for clinical purposes – integration is unsolved!

Private Health vault data
Electronic health record data
Physiological data
Laboratory results

Exposome
Environmental data
Air pollution
Exposure (toxicants)

Collective data
Social data
Fitness, Wellness data
Ambient Assisted Living data
(Non-medical) personal data

Metabolomics
Chemical processes
Cellular reactions
Enzymatic reactions

Foodomics, Lipidomics
Nutrition data (Nutrigenomics)
Diet data (allergenics)

Metabolomics
Chemical processes
Cellular reactions
Enzymatic reactions

Imaging data
X-Ray, ultrasound, MR, CT, PET, cams, observation (e.g. sleep laboratory), gait (child walking)

Proteomics
Protein-Protein Interactions

Epigenetics
Epigenetic modifications

Genomics

Transcriptomics
RNA, mRNA, rRNA, tRNA

Slide 2-2: Taxonomy of data

- **Physical level** -> bit = binary digit = **b**asic **i**ndissolube uni**t** (= Shannon, Sh), ≠ Bit (!) in Quantum Systems -> qubit
- **Logical Level** -> integers, booleans, characters, floating-point numbers, alphanumeric strings, …
- **Conceptual (Abstract) Level** -> data-structures, e.g. lists, arrays, trees, graphs, …
- **Technical Level** -> Application data, e.g. text, graphics, images, audio, video, multimedia, …
- **“Hospital Level”** -> Narrative (textual) data, genetic data, numerical measurements (physiological data, lab results, vital signs, …), recorded signals (ECG, EEG, …), Images (cams, x-ray, MR, CT, PET, …)

Examples: Imaging Data

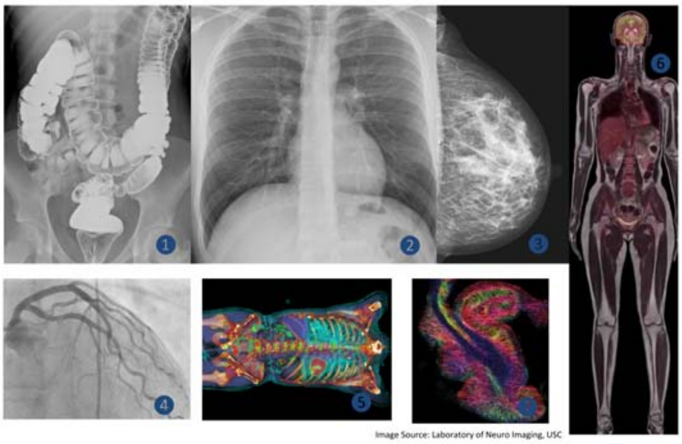
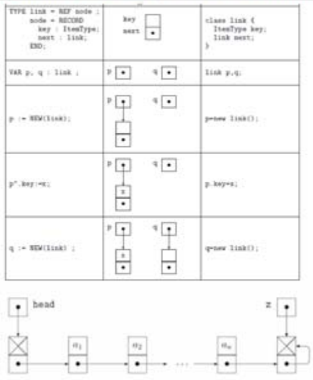


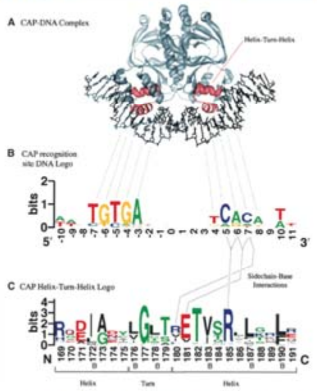
Image Source: Laboratory of Neuro Imaging, USC

Slide 2-3: Example Data Structures (1/3): List



```
TYPE link = REF node ;
     node = RECORD
       key : ItemType;
       next : link;
     END;
VAR p, q : link ;
p := NEW(link);
p^.key:=x;
q := NEW(link) ;
```

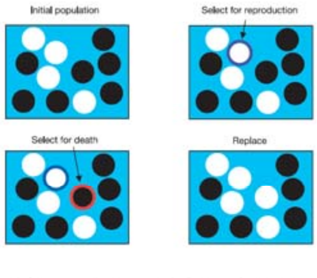
```
class link {
  ItemType key;
  link next;
}
link p,q;
p=new link();
p.key=x;
q=new link();
```

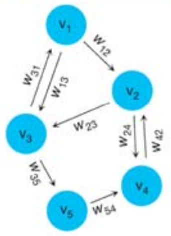

A CAP-DNA Complex
Helix-Turn-Helix
B CAP recognition site DNA Logo
bits
2
1
0
5'
3'
TGTGA
TCACA
C CAP Helix-Turn-Helix Logo
Sidechain-Base Interactions
N
C
Helix
Turn
Helix

Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. (2004) WebLogo: A sequence logo generator. Genome Research, 14, 6, 1188-1190.

Slide 2-4: Example Data Structures (2/3): Graph

Evolutionary dynamics act on populations.
Neither genes, nor cells, nor individuals evolve;
only populations evolve.



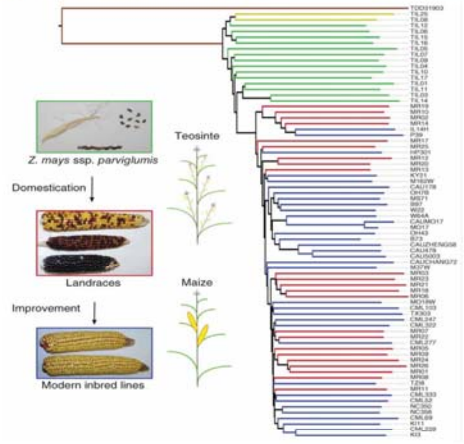


$$W=\begin{bmatrix} 0 & w_{12} & w_{13} & 0 & 0 \\ 0 & 0 & w_{23} & w_{24} & 0 \\ w_{31} & 0 & 0 & 0 & w_{35} \\ 0 & w_{42} & 0 & 0 & 0 \\ 0 & 0 & 0 & w_{54} & 0 \end{bmatrix}$$

Lieberman, E., Hauert, C. & Nowak, M. A. (2005) Evolutionary dynamics on graphs. *Nature, 433, 7023, 312-316.*

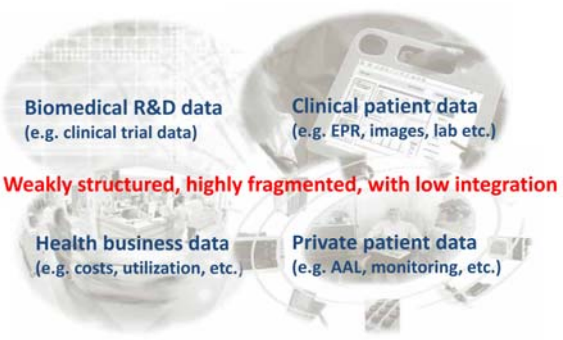
Slide 2-5: Example Data Structures (3/3) Tree

Hufford et. al. 2012. Comparative population genomics of maize domestication and improvement. *Nature Genetics, 44, (7), 808-811.*



Data Integration and Data Fusion in the Life Sciences

Slide 2-6: "Big Data" pools in the health domain



Biomedical R&D data
(e.g. clinical trial data)

Clinical patient data
(e.g. EPR, images, lab etc.)

Weakly structured, highly fragmented, with low integration

Health business data
(e.g. costs, utilization, etc.)

Private patient data
(e.g. AAL, monitoring, etc.)

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. & Byers, A. H. (2011) *Big data: The next frontier for innovation, competition, and productivity. Washington (DC), McKinsey Global Institute.*

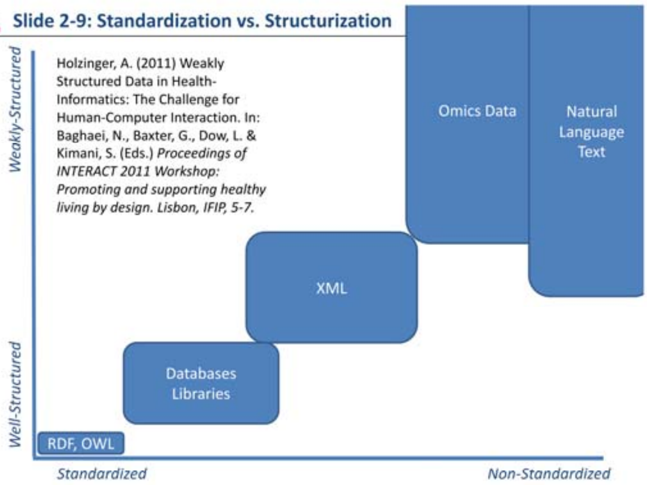
Slide 2-7a: Omics-data integration (1/2)

Genomics	Transcriptomics	Proteomics	Metabolomics	Protein–DNA interactions	Protein–protein interactions	Fluxomics	Phenomics
Genomics (sequence annotation)	• ORF validation • Regulatory element identification	• SNP effect on protein activity or abundance	• Enzyme annotation	• Binding-site identification	• Functional annotation	• Functional annotation	• Functional annotation • Biomarkers
	Transcriptomics (microarray, SAGE)	• Protein: transcript correlation	• Enzyme annotation	• Gene-regulatory networks	• Functional annotation • Protein complex identification		• Functional annotation
		Proteomics (abundance, post-translational modification)	• Enzyme annotation	• Regulatory complex identification	• Differential complex formation	• Enzyme capacity	• Functional annotation
			Metabolomics (metabolite abundance)	• Metabolic-transcriptional response		• Metabolic pathway bottlenecks	• Metabolic flexibility • Metabolic engineering
				Protein–DNA interactions (ChIP–chip)	• Signalling cascades		• Dynamic network responses
					Protein–protein interactions (yeast 2H, coAP–MS)		• Pathway identification activity
						Fluxomics (isotopic tracing)	• Metabolic engineering
							Phenomics (phenotype arrays, RNAi screens, synthetic lethals)

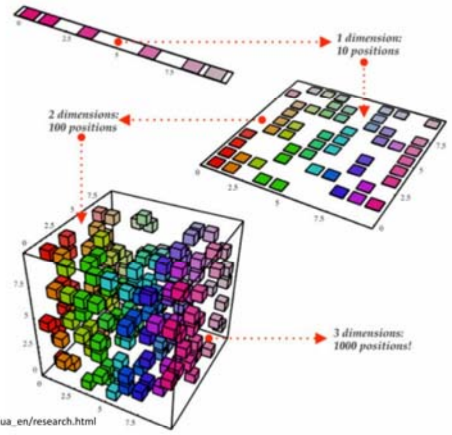
Joyce, A. R. & Palsson, B. Ø. 2006. The model organism as a system: integrating'omics' data sets. *Nature Reviews Molecular Cell Biology, 7,* ***198-210.***

Slide 2-9: Standardization vs. Structurization

Holzinger, A. (2011) Weakly Structured Data in Health-Informatics: The Challenge for Human-Computer Interaction. In: Baghaei, N., Baxter, G., Dow, L. & Kimani, S. (Eds.) *Proceedings of INTERACT 2011 Workshop: Promoting and supporting healthy living by design. Lisbon, IFIP, 5-7.*



Note: The curse of dimensionality


1 dimension: 10 positions
2 dimensions: 100 positions
3 dimensions: 1000 positions!

Bengio, S. & Bengio, Y. 2000. Taking on the curse of dimensionality in joint distributions using neural networks. IEEE Transactions on Neural Networks, 11, (3), 550-557.

http://www.iro.umontreal.ca/~bengioy/yoshua_en/research.html

Slide 2-10: Data Dimensionality examples

- 0-D data = a data point existing isolated from other data, e.g. integers, letters, Booleans, etc.
- 1-D data = consist of a string of 0-D data, e.g. Sequences representing nucleotide bases and amino acids, SMILES etc.
- 2-D data = having spatial component, such as images, NMR-spectra etc.
- 2.5-D data = can be stored as a 2-D matrix, but can represent biological entities in three or more dimensions, e.g. PDB records
- 3-D data = having 3-D spatial component, e.g. image voxels, e-density maps, etc.
- H-D Data = data having arbitrarily high dimensions

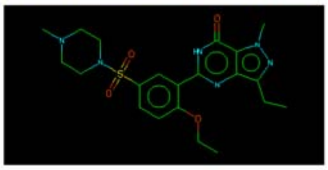
Example: 1-D data (univariate sequential data objects)

SMILES (Simplified Molecular Input Line Entry Specification)

... is a compact machine and human-readable chemical nomenclature:

e.g. Viagra:

CCc1nn(C)c2c(=O)[nH]c(nc12)c3cc(ccc3OCC)S(=O)(=O)N4CCN(C)CC4



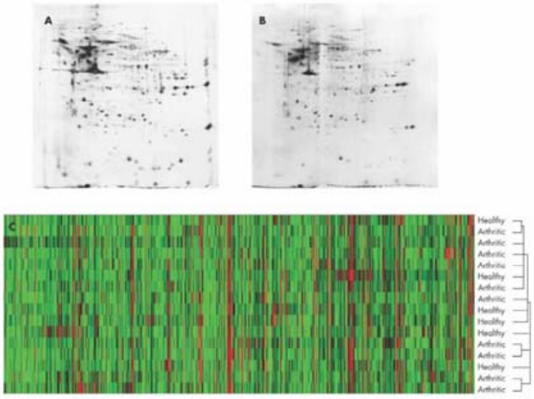
...is Canonicalizable

...is Comprehensive

...is Well Documented

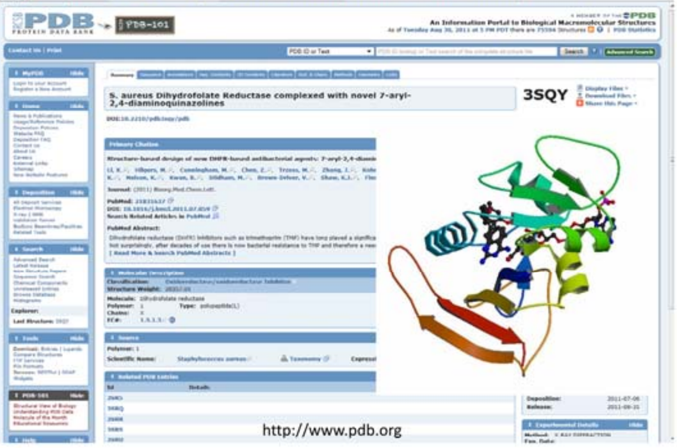
http://www.daylight.com/dayhtml_tutorials/languages/smiles/index.html

Example: 2-D data (bivariate data)



Kastrinaki et al. (2008) Functional, molecular & proteomic characterisation of bone marrow mesenchymal stem cells in rheumatoid arthritis. *Annals of Rheumatic Diseases, 67, 6, 741-749.*

Example: 2.5-D data (structural information and metadata)



PDB PROTEIN DATA BANK

PDB-101

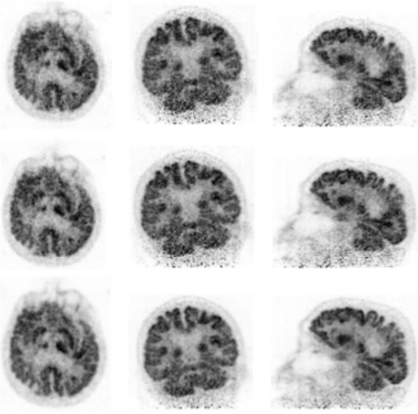
An Information Portal to Biological Macromolecular Structures

S. aureus Dihydrofolate Reductase complexed with novel 7-aryl-2,4-diaminoquinazolines

3SQY

http://www.pdb.org

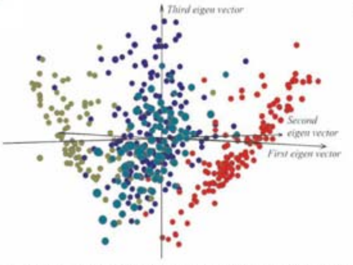
Example: 3-D Voxel data (volumetric picture elements)



Scheins, J. J., Herzog, H. & Shah, N. J. (2011) Fully-3D PET Image Reconstruction Using Scanner-Independent, Adaptive Projection Data and Highly Rotation-Symmetric Voxel Assemblies. *Medical Imaging, IEEE Transactions on, 30, 3, 879-892.*

Slide 2-11 A space is a set of points

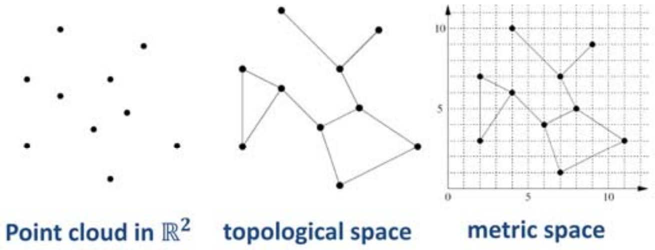
$$f : X \to \mathbb{R}$$



Hou, J., Sims, G. E., Zhang, C. & Kim, S.-H. 2003. A global representation of the protein fold space. *Proceedings of the National Academy of Sciences, 100, (5), 2386-2390.*

Slide 2-12 Point Cloud Data Sets

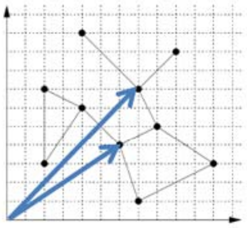
Let us collect n-dimensional i observations: $\boldsymbol{x}_i = [x_{i1}, \dots, x_{in}]$



Zomorodian, A. J. 2005. *Topology for computing, Cambridge (MA), Cambridge University Press.*

Slide 2-13: Example Metric Space

A set S with a metric function d is a metric space



$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$$

Doob, J. L. 1994. *Measure theory, Springer New York.*

Slide 2-14 Mapping Data $\mathbb{R}^n + t \to \mathbb{R}^2 + t \text{ or } \mathbb{R}^3 + t$

Cognitive Space

Computational space

Perception

Visualization

$\mathbb{R}^2$
$\mathbb{R}^n$

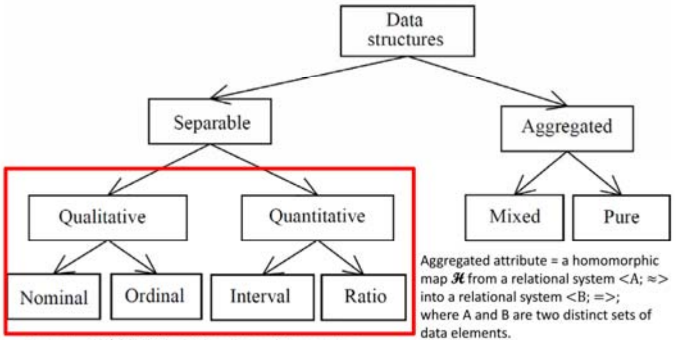
Human intelligence

Machine intelligence

Human **I**nteraction **C**omputer

Holzinger, A. 2012. On Knowledge Discovery and interactive intelligent visualization of biomedical data. In: DATA - International Conference on Data Technologies and Applications.

Example: Data structures - Classification



Aggregated attribute = a homomorphic map $\mathcal{H}$ from a relational system <A; ≈> into a relational system <B; =>; where A and B are two distinct sets of data elements.
This is in contrast with other attributes since the set B is the set of data elements instead of atomic values.

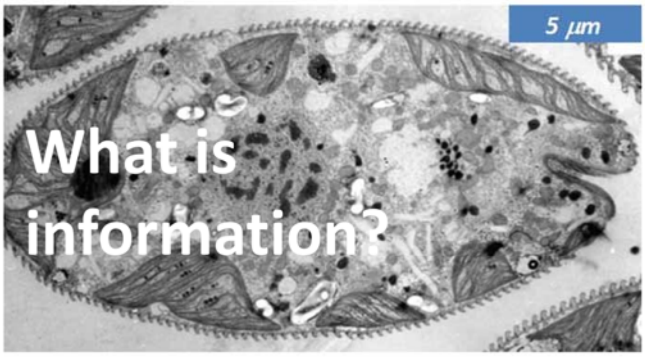
Dastani, M. (2002) The Role of Visual Perception in Data Visualization. *Journal of Visual Languages and Computing, 13, 601-622.*

Slide 2-15: Categorization of Data (Classic "scales")

Scale	Empirical Operation	Mathem. Group Structure	Transf. in $\mathbb{R}$	Basic Statistics	Mathematical Operations
NOMINAL	Determination of equality	Permutation $x' = f(x)$ x … 1-to-1	$x \mapsto f(x)$	Mode, contingency correlation	=, ≠
ORDINAL	Determination of more/less	Isotonic $x' = f(x)$ x … mono-tonic incr.	$x \mapsto f(x)$	Median, Percentiles	=, ≠, >, <
INTERVAL	Determination of equality of intervals or differences	General linear $x' = ax + b$	$x \mapsto rx+s$	Mean, Std.Dev. Rank-Order Corr., Prod.-Moment Corr.	=, ≠, >, <, -, +
RATIO	Determination of equality or ratios	Similarity $x' = ax$	$x \mapsto rx$	Coefficient of variation	=, ≠, >, <, -, +, *, ÷

Stevens, S. S. (1946) On the theory of scales of measurement. *Science, 103, 677-680.*

Slide 2-18: Life is complex information

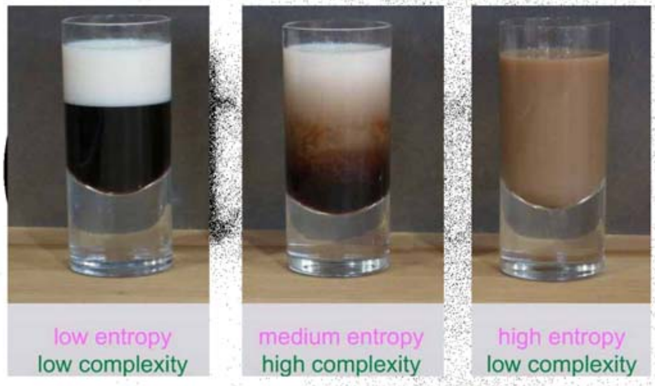


Lane, N. & Martin, W. (2010) The energetics of genome complexity. *Nature, 467, 7318, 929-934.*

Probability > Information > Entropy (Quick Preview)

- Information is the reduction of uncertainty
- If something is 100 % certain its uncertainty = 0
- Uncertainty is a max. if all choices are equally probable
- Uncertainty (as information) sums up for independent sources

(Physical) Entropy as measure for disorder



http://www.scottaaronson.com

Physical Entropy

$$\neq$$

Information Entropy

εντροπια

> *My greatest concern was what to call it. I thought of calling it "information", but the word was overly used, so I decided to call it "uncertainty". When I discussed it with John von Neumann, he had a better idea. Von Neumann told me, "You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, nobody knows what entropy really is, so in a debate you will always have the advantage."*

Tribus, M. & McIrvine, E. C. (1971) Energy and Information. *Scientific American, 225, 3, 179-184.*

How to measure Information?

14

- How informative is an observation of a particular value x for a random variable X with the probability $P(X)$?
- If $P(X = x)$ is high → not surprising (not informative)
- If $P(X = x) \approx 0$ → surprising (novel → previously unknown → informative)
- Shannon [1] showed that the best way to quantify the concept of information of an event X = x is to take the inverse of the probability:

$$\log \frac{1}{P(X = x)} = -\log P(X = x)$$

[1] Shannon, C. E. & Weaver, W. 1949. The Mathematical Theory of Communication, Urbana, Univ. of Illinois Press.

Information (Shannon!) Entropy

15

Shannon called $H(X)$ the entropy of X and used it as a measure of the randomness (= uncertainty) of the distribution $P(X)$:

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log_b P(x_i)$$

for $b = 10$ we call the unit "Hartley" (decimal digit)
for $b = 2$ we call the unit "bit" (binary digit)
for $b = e$ we call the unit "nat" (natural digit)
If we have instead of a discrete source a continuous signal, the sum can be replaced by the integral:

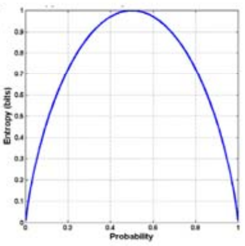
$$H(X) = -\int P(x) \log_b P(x) dx$$

Original Work: Q … discrete data source!

$$Q \ \dots P = \{p_1, \dots, p_n\} \qquad H(Q) = -\sum_{i=1}^{n}(p_i * \log p_i)$$

$$Qb = \{a_1, a_2\} \textit{ with } P = \{p, 1-p\}$$

$$H\,(Qb) = p * \log\frac{1}{p} + p * \log\frac{1}{1-p}$$



Shannon, C. E. (1948) A Mathematical Theory of Communication. *Bell System Technical Journal, 27, 379-423.*

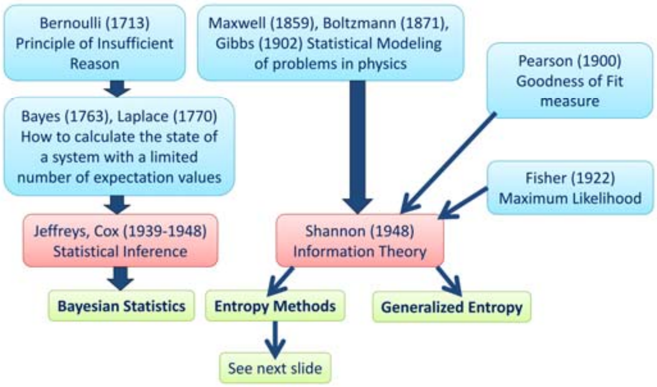
Shannon, C. E. & Weaver, W. (1949) *The Mathematical Theory of Communication. Urbana (IL), University of Illinois Press.*

Entropic methods – why and what for?

- 1) Set of noisy, complex, time series data
- 2) Extract information out of the data
- 3) to support a previous set hypothesis
- Information + Statistics + Inference
- = powerful methods for many sciences
- Application e.g. in biomedical informatics for analysis of ECG, MRI, CT, PET, sequences and proteins, DNA, topography, for modeling etc. etc.

Mayer, C., Bachler, M., Hortenhuber, M., Stocker, C., Holzinger, A. & Wassertheurer, S. 2014. Selection of entropy-measure parameters for knowledge discovery in heart rate variability data. BMC Bioinformatics, 15, (Suppl 6), S2.

An overview on the History of Entropy



Bernoulli (1713) Principle of Insufficient Reason

Maxwell (1859), Boltzmann (1871), Gibbs (1902) Statistical Modeling of problems in physics

Pearson (1900) Goodness of Fit measure

Bayes (1763), Laplace (1770) How to calculate the state of a system with a limited number of expectation values

Fisher (1922) Maximum Likelihood

Jeffreys, Cox (1939-1948) Statistical Inference

Shannon (1948) Information Theory

Bayesian Statistics

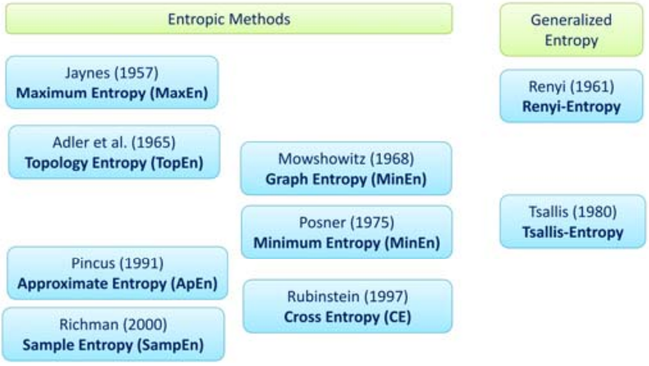
Entropy Methods

Generalized Entropy

See next slide

confer also with: Golan, A. (2008) Information and Entropy Econometric: A Review and Synthesis. *Foundations and Trends in Econometrics, 2, 1-2, 1-145.*

Towards a Taxonomy of Entropic Methods



Entropic Methods

Generalized Entropy

Jaynes (1957) **Maximum Entropy (MaxEn)**

Renyi (1961) **Renyi-Entropy**

Adler et al. (1965) **Topology Entropy (TopEn)**

Mowshowitz (1968) **Graph Entropy (MinEn)**

Tsallis (1980) **Tsallis-Entropy**

Posner (1975) **Minimum Entropy (MinEn)**

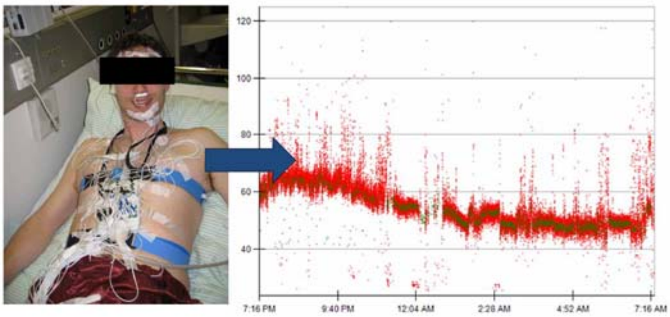
Pincus (1991) **Approximate Entropy (ApEn)**

Rubinstein (1997) **Cross Entropy (CE)**

Richman (2000) **Sample Entropy (SampEn)**

Holzinger, A., Hörtenhuber, M., Mayer, C., Bachler, M., Wassertheurer, S., Pinho, A. & Koslicki, D. 2014. On Entropy-Based Data Mining. In: Holzinger, A. & Jurisica, I. (eds.) Lecture Notes in Computer Science, LNCS 8401. Berlin Heidelberg: Springer, pp. 209-226.

Example of the usefulness of ApEn (1/3)



Holzinger, A., Stocker, C., Bruschi, M., Auinger, A., Silva, H., Gamboa, H. & Fred, A. 2012. On Applying Approximate Entropy to ECG Signals for Knowledge Discovery on the Example of Big Sensor Data. *In: Huang, R., Ghorbani, A., Pasi, G., Yamaguchi, T., Yen, N. & Jin, B. (eds.) Active Media Technology, Lecture Notes in Computer Science, LNCS 7669. Berlin Heidelberg: Springer, pp. 646-657.*

EU Project EMERGE (2007-2010)

Example of the usefulness of ApEn (2/3)

$$Let\colon\ \langle x_n \rangle = \{x_1, x_2, \dots, x_N\}$$

$$\vec{X}_i = (x_i, x_{(i+1)}, \dots, x_{(i+m-1)})$$

$$\|\vec{X}_i, \vec{X}_j\| = \max_{k=1,2,\dots,m}\left(\left|x_{(i+k-1)} - x_{(j+k-1)}\right|\right)$$

$$\boxed{\widetilde{\mathrm{H}}(m,r) = \lim_{N\to\infty}\left[\phi^m(r) - \phi^{m+1}(r)\right]}$$

$$C_r^m(i) = \frac{N^m(i)}{N-m+1} \qquad \phi^m(r) = \frac{1}{N-m+1}\sum_{t=1}^{N-m+1}\ln C_r^m(i)$$

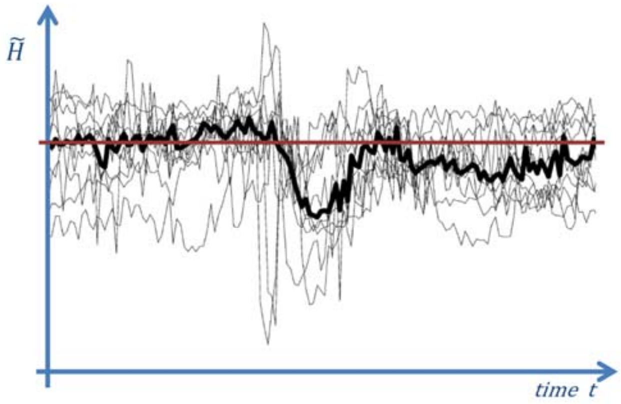
Pincus, S. M. (1991) Approximate Entropy as a measure of system complexity. *Proceedings of the National Academy of Sciences of the United States of America, 88, 6, 2297-2301.*

Algorithm – Pseudo Code – Python Example

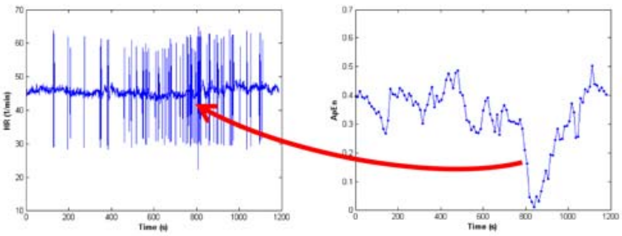
- 1: Sample N raw data from a time series
- 2: Set m (integer), r (real)
- 3: Build vectors in $\mathbb{R}^m$
- 4: Measure the distance between every component, i.e. the max. difference between the scalar components
- 5: Define the start parameters (constants)
- 6: Calculate the ApEn (H)

```
import numpy as np
def ApEn(U, m, r):
   def _maxdist(x_i, x_j):
        return max([abs(ua - va) for ua, va in zip(x_i, x_j)])
   def _phi(m):
 x = [[U[j] for j in range(i, i + m - 1 + 1)] for i in range(N - m + 1)]
 C = [len([1 for x_j in x if _maxdist(x_i, x_j) <= r]) / (N - m + 1.0) for x_i in x]
        return (N - m + 1.0)**(-1) * sum(np.log(C))
    N = len(U)
    return abs(_phi(m + 1) - _phi(m))
print ApEn(U, 2, 3)
```

Example: ApEn (2)

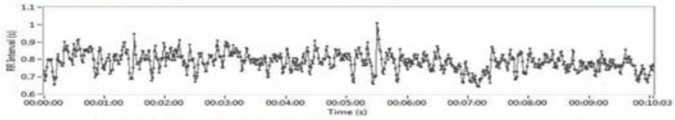


ApEn



Holzinger, A., Hörtenhuber, M., Mayer, C., Bachler, M., Wassertheurer, S., Pinho, A. & Koslicki, D. 2014. On Entropy-Based Data Mining. In: Holzinger, A. & Jurisica, I. (eds.) Interactive Knowledge Discovery and Data Mining in Biomedical Informatics, Lecture Notes in Computer Science, LNCS 8401. Berlin Heidelberg: Springer, pp. 209-226.

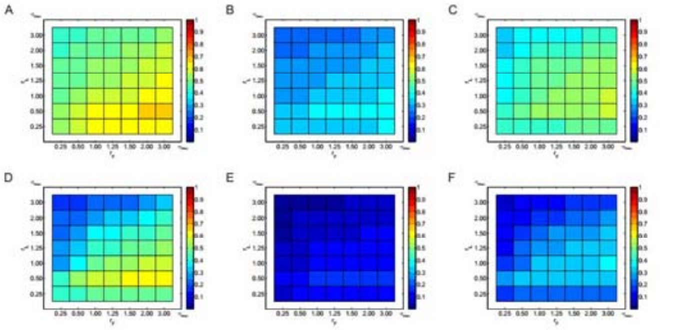
Summary: Example Heart Rate Variability



- Heart Rate Variability (HRV) can be used as a marker of cardiovascular health status.
- Entropy measures represent a family of new methods to quantify the variability of the heart rate.
- Promising approach, due to ability to discover certain patterns and shifts in the "apparent ensemble amount of randomness" of stochastic processes,
- measure randomness and **predictability of processes.**

Mayer, C., Bachler, M., Holzinger, A., Stein, P. K. & Wassertheurer, S. 2016. The Effect of Threshold Values and Weighting Factors on the Association between Entropy Measures and Mortality after Myocardial Infarction in the Cardiac Arrhythmia Suppression Trial (CAST). Entropy, 18, (4), 129, doi::10.3390/e18040129.

Baseline: A,B,C – After treatment: D,E,F (N=1200)



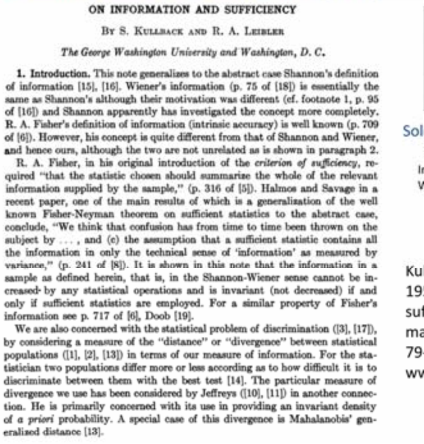
Mayer, C., Bachler, M., Holzinger, A., Stein, P. K. & Wassertheurer, S. 2016. The Effect of Threshold Values and Weighting Factors on the Association between Entropy Measures and Mortality after Myocardial Infarction in the Cardiac Arrhythmia Suppression Trial (CAST). Entropy, 18, (4), 129, doi::10.3390/e18040129.

Mutual Information Cross-Entropy Kullback-Leibler Divergence

Entropy – KL-Div. – Mutual Information

- Entropy:
  - Measure for the **uncertainty** of random variables
- Mutual Information/:
  - measuring the **correlation** of two random variables
- Kullback-Leibler divergence:
  - **comparing two distributions**

Solomon Kullback & Richard Leibler (1951)



ON INFORMATION AND SUFFICIENCY

By S. Kullback and R. A. Leibler

The George Washington University and Washington, D. C.

1. Introduction. This note generalizes to the abstract case Shannon's definition of information [15], [16]. Wiener's information (p. 75 of [18]) is essentially the same as Shannon's although their motivation was different (cf. footnote 1, p. 95 of [16]) and Shannon apparently has investigated the concept more completely. R. A. Fisher's definition of information (intrinsic accuracy) is well known (p. 709 of [6]). However, his concept is quite different from that of Shannon and Wiener, and hence ours, although the two are not unrelated as is shown in paragraph 2.

R. A. Fisher, in his original introduction of the *criterion of sufficiency*, required "that the statistic chosen should summarize the whole of the relevant information supplied by the sample," (p. 316 of [5]). Halmos and Savage in a recent paper, one of the main results of which is a generalization of the well known Fisher-Neyman theorem on sufficient statistics to the abstract case, conclude, "We think that confusion has from time to time been thrown on the subject by . . . , and (c) the assumption that a sufficient statistic contains all the information in only the technical sense of 'information' as measured by variance," (p. 241 of [8]). It is shown in this note that the information in a sample as defined herein, that is, in the Shannon-Wiener sense cannot be increased by any statistical operations and is invariant (not decreased) if and only if sufficient statistics are employed. For a similar property of Fisher's information see p. 717 of [6], Doob [19].

We are also concerned with the statistical problem of discrimination ([3], [17]), by considering a measure of the "distance" or "divergence" between statistical populations ([1], [2], [13]) in terms of our measure of information. For the statistician two populations differ more or less according as to how difficult it is to discriminate between them with the best test [14]. The particular measure of divergence we use has been considered by Jeffreys ([10], [11]) in another connection. He is primarily concerned with its use in providing an invariant density of *a priori* probability. A special case of this divergence is Mahalanobis' generalized distance [13].

Solomon Kullback 1907-1994

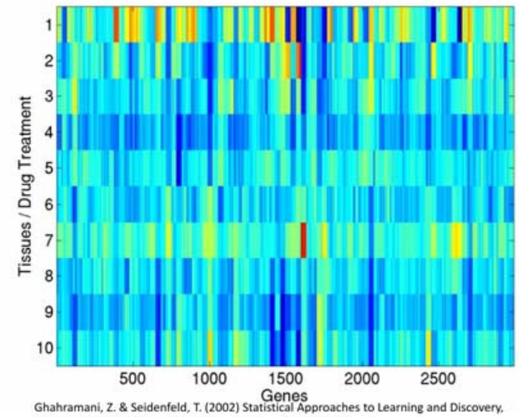
Image Source: Wikipedia

Richard Leibler 1914-2003

Image Source: http://www.datavortex.com/mathematical_heroes/

Kullback, S. & Leibler, R. A. 1951. On information and sufficiency. The annals of mathematical statistics, 22, (1), 79-86, www.jstor.org/stable/2236703

Example: Looking for repetitions (= low entropy)



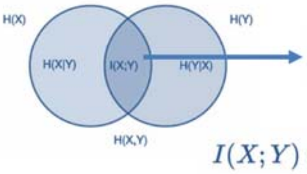
Ghahramani, Z. & Seidenfeld, T. (2002) Statistical Approaches to Learning and Discovery, CALD Carnegie Mellon University, Lecture 1: Information Theory

Conclusion 1) Information Entropy

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log_b P(x_i)$$

- Measuring of uncertainty, complexity, randomness, surprise, …, information
- coding theory
- statistical physics
- handwriting recognition
- machine learning
- etc. etc.

Mutual Information - Conditional Entropy



$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log\left(\frac{p(x,y)}{p(x)\,p(y)}\right)$$

$$I(X;Y) = H(X) - H(X|Y)$$

- In ML we need often to measure the **difference between two probability distributions**

Kullback-Leibler Divergence - discriminative information

For discrete distributions $D_{\mathrm{KL}}(P\|Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$

For continuous distributions $D_{\mathrm{KL}}(P\|Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} \, \mathrm{d}x$

$$\mathrm{KL}(p\|q) \geqslant 0 \qquad \mathrm{KL}(p\|q) \not\equiv \mathrm{KL}(q\|p)$$

KL-divergence can also be used to measure the **distance between two distributions**

Kullback, S. & Leibler, R. A. 1951. On information and sufficiency. The annals of mathematical statistics, 22, (1), 79-86, doi:http://www.jstor.org/stable/2236703

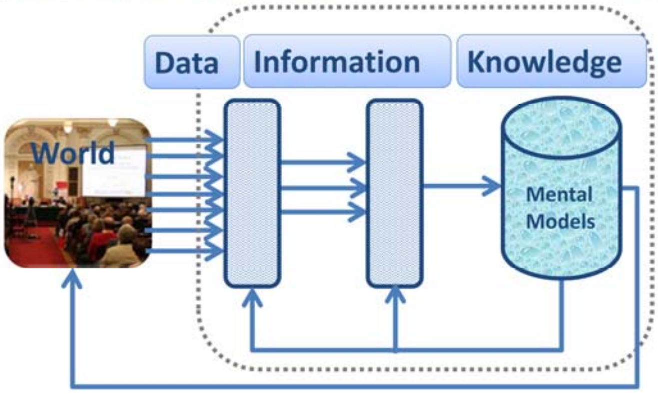
Entropy measures generally …

- … are **robust** against noise;
- … can be applied to **complex time series** with good replication;
- … is **finite** for stochastic, noisy, composite processes – good for automated classification
- … the values correspond directly to irregularities – good for detecting **anomalies**

Knowledge

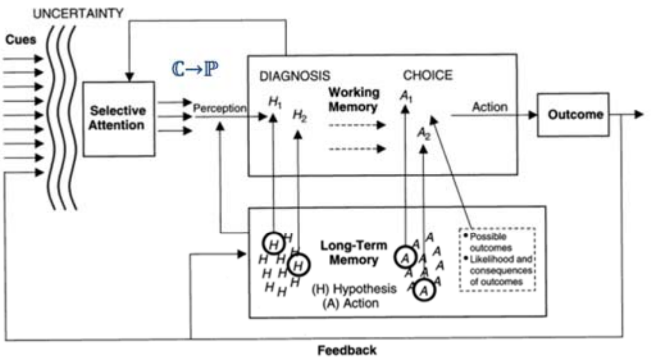
What is Knowledge?

Adaptive Agent in a physical world



Knowledge := a set of expectations

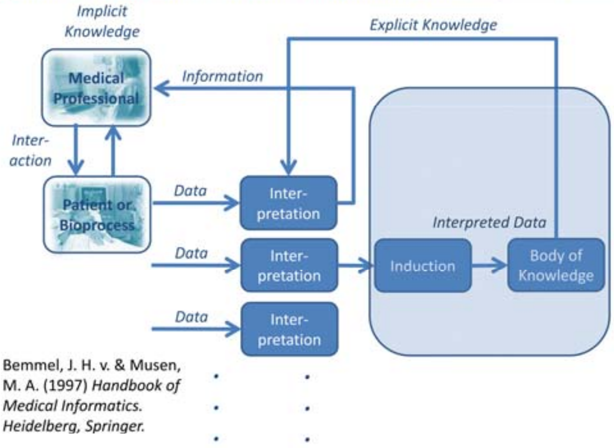
Slide 2-19: Human Information Processing Model



Wickens, C. D. (1984) *Engineering psychology and human performance. Columbus: Merrill.*

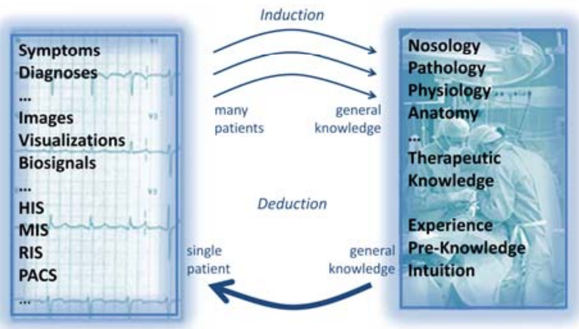
A clinical view on data – information - knowledge

Slide 2-16: Clinical View of Data, Information, Knowledge



Bemmel, J. H. v. & Musen, M. A. (1997) *Handbook of Medical Informatics. Heidelberg, Springer.*

Slide 2-17: From Patient Data to Medical Knowledge

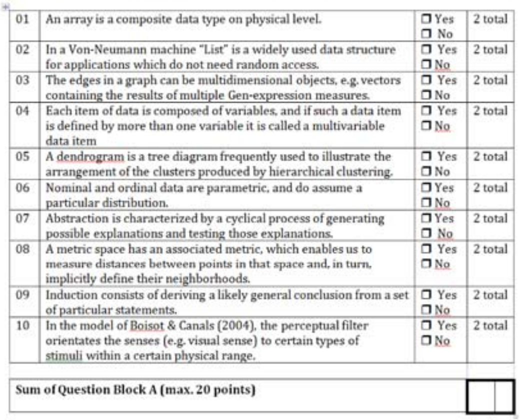


Holzinger (2007)

Thank you!

Questions

Sample Exam Questions – Yes/No Answers



01	An array is a composite data type on physical level.	❐ Yes ❐ No	2 total
02	In a Von-Neumann machine "List" is a widely used data structure for applications which do not need random access.	❐ Yes ❐ No	2 total
03	The edges in a graph can be multidimensional objects, e.g. vectors containing the results of multiple Gen-expression measures.	❐ Yes ❐ No	2 total
04	Each item of data is composed of variables, and if such a data item is defined by more than one variable it is called a multivariable data item	❐ Yes ❐ No	2 total
05	A dendrogram is a tree diagram frequently used to illustrate the arrangement of the clusters produced by hierarchical clustering.	❐ Yes ❐ No	2 total
06	Nominal and ordinal data are parametric, and do assume a particular distribution.	❐ Yes ❐ No	2 total
07	Abstraction is characterized by a cyclical process of generating possible explanations and testing those explanations.	❐ Yes ❐ No	2 total
08	A metric space has an associated metric, which enables us to measure distances between points in that space and, in turn, implicitly define their neighborhoods.	❐ Yes ❐ No	2 total
09	Induction consists of deriving a likely general conclusion from a set of particular statements.	❐ Yes ❐ No	2 total
10	In the model of Boisot & Canals (2004), the perceptual filter orientates the senses (e.g. visual sense) to certain types of stimuli within a certain physical range.	❐ Yes ❐ No	2 total

Sum of Question Block A (max. 20 points) □□

Sample Questions (1)

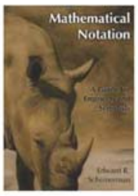
- Why is modeling of artifacts a huge problem?
- What do we need to transfer information into Knowledge?
- What type of data does the PDB basically store?
- What is the "curse of dimensionality"?
- What type of separable data is blood sedimentation rate?
- Is the mathematical operation "multiplication" allowed with ordinal data?
- What characterizes standardized data?
- Why are structural homologies interesting?
- How did Bemmel & van Musen describe the clinical view on data, information and knowledge?
- Where are the differences between patient data and medical knowledge from a clinical viewpoint?
- Which weaknesses of the DIKW Model do you recognize?
- How do we get theories?
- What is the main limitation of transferring data from the computational space into the perceptual space from the viewpoint of the human information processing model?

Sample Questions (2)

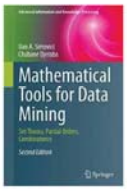
- Why is the knowledge about human information processing necessary for medical informatics?
- What advantages does the Kullback-Leibler divergence offer – what is the drawback?
- What does information interaction mean?
- How does knowledge-assisted visualization work in principle?
- Why is non-structured data an rather incorrect term?
- Give an example of the data structure tree in biomedical informatics!
- Why is data quality important? What are the related issues?
- How do you ensure data accessibility?
- What is the main idea of Shannon's Entropy?
- Why is Entropy interesting for medical informatics?
- What are typical entropic methods?
- What is the main purpose of Approximate Entropy?
- What is the big advantage of entropic methods?
- What are the differences of ApEn and SampEn?
- Which possibilities do you have with Graph Entropy Measures?

Appendix

Recommended Books



Scheinerman, E. R. 2011. Mathematical Notation: A Guide for Engineers and Scientists, Baltimore (MD), Scheinerman.



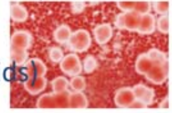
Simovici, D. A. & Djeraba, C. 2014. Mathematical tools for data mining. Second Edition, doi:10.1007/978-1-4471-6407-4.

Rosen, K. H. & Krithivasan, K. 2012. Discrete mathematics and its applications. Seventh Edition, New York, McGraw-Hill.

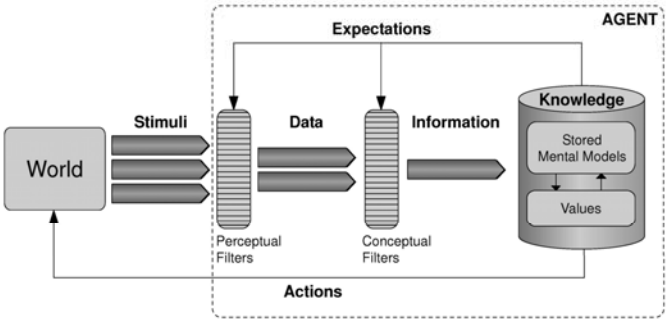
Duda, R. O., Hart, P. E. & Stork, D. G. 2000. Pattern Classification. Second Edition, New York et al., Wiley.

Slide 2-7b: -Omics-data integration (2/2)

- **Genomics** (sequence annotation)
- **Transcriptomics** (microarray)
- **Proteomics** (Proteome Databases)
- **Metabolomics** (enzyme annotation)
- **Fluxomics** (isotopic tracing, metabolic pathways)
- **Phenomics** (biomarkers)
- **Epigenomics** (epigenetic modifications)
- **Microbiomics** (microorganisms)
- **Lipidomics** (pathways of cellular lipids)



Slide 2-20: Knowledge as a set of expectations



Boisot, M. & Canals, A. 2004. Data, information and knowledge: have we got it right? *Journal of Evolutionary Economics, 14, (1), 43-67.*

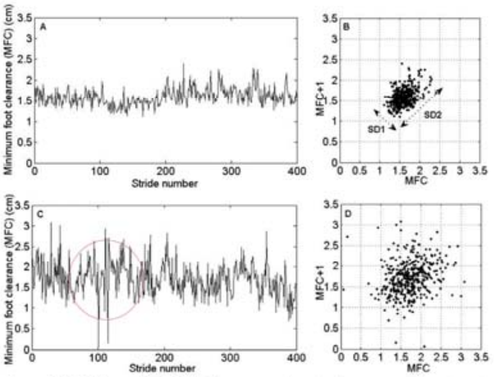
Additional Reading

- **History of Probability Theory**
- Franklin, J. *The Science of Conjecture: Evidence and Probability Before Pascal.* John Hopkins University Press, 2001.
- Jaynes, E. T. *Probability Theory: The Logic of Science.* Cambridge University Press, 2003.
- **Probabilistic Reasoning**
- Gigerenzer, G., and D. J. Murray. *Cognition as Intuitive Statistics.* Hillsdale, NJ: Erlbaum, 1987.
- Gilovich, T., D. Griffin, and D. Kahneman, eds. *Heuristics and Biases: The Psychology of Intuitive Judgment.* Cambridge University Press, 2002.
- Kahneman, D., P. Slovic, and A. Tversky, eds. *Judgment under Uncertainty: Heuristics and Biases.* Cambridge University Press, 1982.
- **Bayesian Networks**
- Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufman, San Mateo, CA, 1988.
- Breese, J. S. "Construction of Belief and Decision Networks." *Computational Intelligence* 8, 4 (1992): 624–647.
- F. Bacchus, A. J. Grove, J. Y. Halpern, and D. Koller. "Statistical Foundations for Default Reasoning." *Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI).* Chambery, France, August 1993, pp. 563-569.
- **Multiple-Instance Bayesian Networks**
- Pasula, H., and S. Russell. "Approximate Inference for First-order Probabilistic Languages." *IJCAI-01.* Seattle, WA, 2001, pp. 741–748.
- Halpern, J. Y. "An Analysis of First-order Logics of Probability." *Artificial Intelligence* 46, 3 (1990): 311–350.
- D. Koller, and A. Pfeffer. "Object-Oriented Bayesian Networks." *Proceedings of the 13th Annual Conference on Uncertainty in AI (UAI).* Providence, Rhode Island, 1997, pp. 302-313.

Solutions to the warm-up quiz

- There a gap (ocean) between these two worlds due to the inherent complexity of the fields with different goals and tasks.
- Uncertainty is one of the four main problems, among heterogeneity, dimensionality and complexity – and this inherently to our real-world: we are surrounded by vague, imprecise, uncertain information
- The posterior can be calculated as the likelihood times the prior through the evidence –this so cool because the inverse probability allows us to learn from data and to make predictions
  D=set of data, theta
- Where is the highest certainty in this image – how is now the degree of uncertainty described?
- A best practice example for fully automated ML – autonomous driving
- Medical Decision Making is a search task in arbitrarily high dimensions – problem is limited time
- iML –sometimes we need a human in the loop – e.g. in helping to solve NP-hard problems
- A big challenge is to map the results down into R2 - visualization
- A future trend is in personalized medicine – a step before is stratified medicine

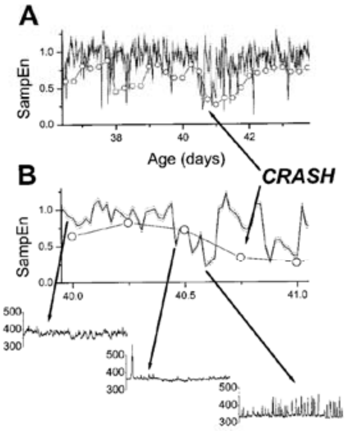
Back-up Slide: Poincare Plot for gait analysis



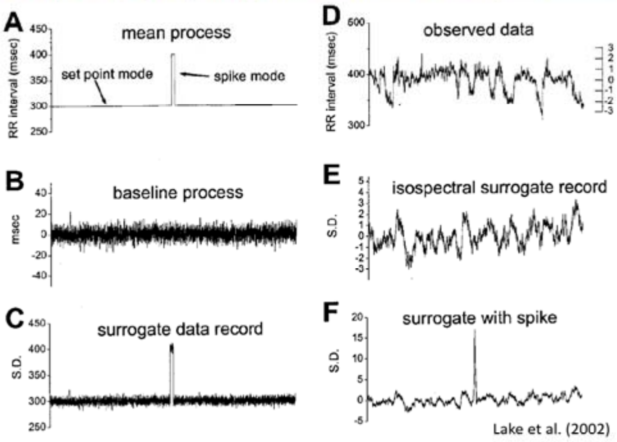
Khandoker, A., Palaniswami, M. & Begg, R. (2008) A comparative study on approximate entropy measure and poincare plot indexes of minimum foot clearance variability in the elderly during walking. *Journal of NeuroEngineering and Rehabilitation, 5, 1, 4.*

Back-up Slide: SampEn (1/2)

Lake, D. E., Richman, J. S., Griffin, M. P. & Moorman, J. R. (2002) Sample entropy analysis of neonatal heart rate variability. *American Journal of Physiology-Regulatory Integrative and Comparative Physiology, 283, **3**, R789-R797.*



Backup: SampEn (2/2) Surrogate data heart rate variability



Lake et al. (2002)

Backup Slide: Comparison ApEn - SampEn

ApEn

Given a signal x(n)=x(1), x(2),..., x(N), where N is the total number of data points, ApEn algorithm can be summarized as follows [1]:

1) Form *m*-vectors, *X(1)* to *X(N-m+1)* defined by:

$$X(i)=[x(i),x(i+1),\dots,X(i+m-1)] \quad i=1,N-m+1 \quad (1)$$

2) Define the distance *d[X(i),X(j)]* between vectors *X(i)* and *X(j)* as the maximum absolute difference between their respective scalar components:

$$d[X(i),X(j)]=\max_{k=0,m-1}[|x(i+k)-x(j+k)|] \quad (2)$$

3) Define for each i, for i=1, N-m+1, let

$$C_r^m(i)=V^m(i)/(N-m+1)$$
$$where \quad V^m(i)=no.of\ d[X(i),X(j)]\le r \quad (3)$$

4) Take the natural logarithm of each $C_r^m(i)$, and average it over *i* as defined in step 3):

$$\phi^m(r)=\frac{1}{N-m+1}\sum_{i=1}^{N-m+1}\ln(C_r^m(i)) \quad (4)$$

5) Increase the dimension to m+1 and repeat steps 1) to 4).

6) Calculate ApEn value for a finite data length of *N*:

$$ApEn(m,r,N)=\phi^m(r)-\phi^{m+1}(r) \quad (5)$$

SampEn

Given a signal x(n)=x(1), x(2),..., x(N), where N is the total number of data points, SampEn algorithm can be summarized as follows [5]:

1) Form *m*-vectors, *X(1)* to *X(N-m+1)* defined by:

$$X(i)=[x(i),x(i+1),\dots,X(i+m-1)] \quad i=1,N-m+1 \quad (6)$$

2) Define the distance $d_m[X(i),X(j)]$ between vectors *X(i)* and *X(j)* as the maximum absolute difference between their respective scalar components:

$$d_m[X(i),X(j)]=\max_{k=0,m-1}[|x(i+k)-x(j+k)|] \quad (7)$$

3) Define for each i, for i=1, N-m, let

$$B_i^m(r)=\frac{1}{N-m-1}\times no.of\ d_m[X(i),X(j)]\le r,\ i\ne j \quad (8)$$

4) Similarly, define for each i, for i=1, N-m, let

$$A_i^m(r)=\frac{1}{N-m-1}\times no.of\ d_{m+1}[X(i),X(j)]\le r, i\ne j \quad (9)$$

5) Define $B^m(r)=\frac{1}{N-m}\sum_{i=1}^{N-m}B_i^m(r)$ (10)

$$A^m(r)=\frac{1}{N-m}\sum_{i=1}^{N-m}A_i^m(r) \quad (11)$$

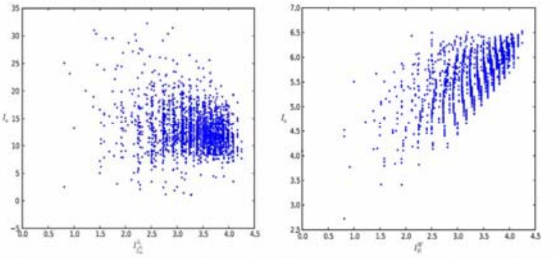
6) SampEn value for a finite data length of *N* can be estimated:

$$SampEn(m,r,N)=-\ln\left(A^m(r)/B^m(r)\right) \quad (12)$$

Xinnian, C. et al. (2005). *Comparison of the Use of Approximate Entropy and Sample Entropy: Applications to Neural Respiratory Signal. Engineering in Medicine and Biology IEEE-EMBS 2005, 4212-4215.*

Backup Slide: Graph Entropy Measures

- The most important question: Which kind of structural information does the entropy measure detect?
- the topological complexity of a molecular graph is characterized by its number of vertices and edges, branching, cyclicity etc.



Dehmer, M. & Mowshowitz, A. (2011) A history of graph entropy measures. *Information Sciences, 181, 1, 57-78.*

Backup: English/German Subject Codes OEFOS 2012

106005	Bioinformatics	Bioinformatik
106007	Biostatistics	Biostatistik
304005	Medical Biotechnology	Medizinische Biotechnologie
305901	Computer-aided diagnosis and therapy	Computerunterstützte Diagnose und Therapie
304003	Genetic engineering, -technology	Gentechnik, -technologie
3906 (old)	Medical computer sciences	Medizinische Computerwissenschaften
305906	Medical cybernetics	Medizinische Kybernetik
305904	Medical documentation	Medizinische Dokumentation
305905	Medical informatics	Medizinische Informatik
305907	Medical statistics	Medizinische Statistik

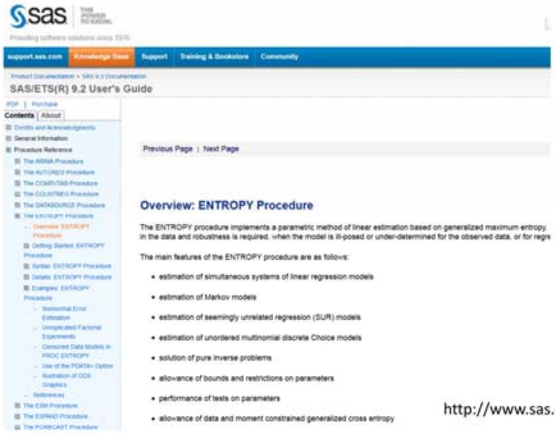
http://www.statistik.at

Backup: English/German Subject Codes OEFOS 2012

102001	Artificial Intelligence	Künstliche Intelligenz
102032	Computational Intelligence	Computational Intelligence
102033	Data Mining	Data Mining
102013	Human-Computer Interaction	Human-Computer Interaction
102014	Information design	Informationsdesign
102015	Information systems	Informationssysteme
102028	Knowledge engineering	Knowledge Engineering
102019	Machine Learning	Maschinelles Lernen
102020	Medical Informatics	Medizinische Informatik
102021	Pervasive Computing	Pervasive Computing
102022	Software development	Softwarenetwicklung
102027	Web engineering	Web Engineering

http://www.statistik.at

Backup Slide: Statistical Analysis Software (SAS)


sas THE POWER TO KNOW.
Providing software solutions since 1976
support.sas.com
Knowledge Base
Support
Training & Bookstore
Community
Product Documentation > SAS 9.2 Documentation
PDF | Purchase
Contents
About
Credits and Acknowledgments
General Information
Procedure Reference
The ARIMA Procedure
The AUTOREG Procedure
The COMPUTAB Procedure
The COUNTREG Procedure
The DATASOURCE Procedure
The ENTROPY Procedure
Overview: ENTROPY Procedure
Getting Started: ENTROPY Procedure
Syntax: ENTROPY Procedure
Details: ENTROPY Procedure
Examples: ENTROPY Procedure
Nonnormal Error Estimation
Unreplicated Factorial Experiments
Censored Data Models in PROC ENTROPY
Use of the PDATA= Option
Illustration of ODS Graphics
References
The ESM Procedure
The EXPAND Procedure
The FORECAST Procedure
Previous Page | Next Page

SAS/ETS(R) 9.2 User's Guide

Overview: ENTROPY Procedure

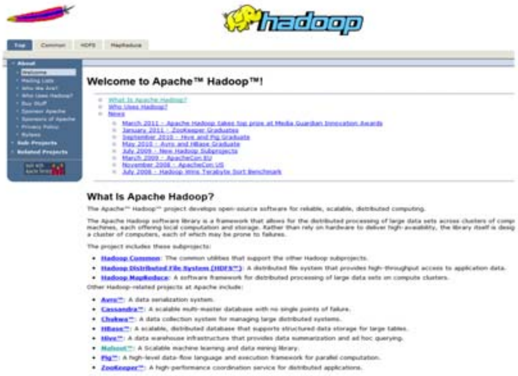
The ENTROPY procedure implements a parametric method of linear estimation based on generalized maximum entropy. in the data and robustness is required, when the model is ill-posed or under-determined for the observed data, or for regre

The main features of the ENTROPY procedure are as follows:

- estimation of simultaneous systems of linear regression models
- estimation of Markov models
- estimation of seemingly unrelated regression (SUR) models
- estimation of unordered multinomial discrete Choice models
- solution of pure inverse problems
- allowance of bounds and restrictions on parameters
- performance of tests on parameters
- allowance of data and moment constrained generalized cross entropy

http://www.sas.com

Backup Slide: Example Tool for large data sets - Hadoop



Welcome to Apache™ Hadoop™!

- What Is Apache Hadoop?
- Who Uses Hadoop?
- News
  - March 2011 - Apache Hadoop takes top prize at Media Guardian Innovation Awards
  - January 2011 - ZooKeeper Graduates
  - September 2010 - Hive and Pig Graduate
  - May 2010 - Avro and HBase Graduate
  - July 2009 - New Hadoop Subprojects
  - March 2009 - ApacheCon EU
  - November 2008 - ApacheCon US
  - July 2008 - Hadoop Wins Terabyte Sort Benchmark

What Is Apache Hadoop?

The Apache™ Hadoop™ project develops open-source software for reliable, scalable, distributed computing.

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of comp machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is desig a cluster of computers, each of which may be prone to failures.

The project includes these subprojects:

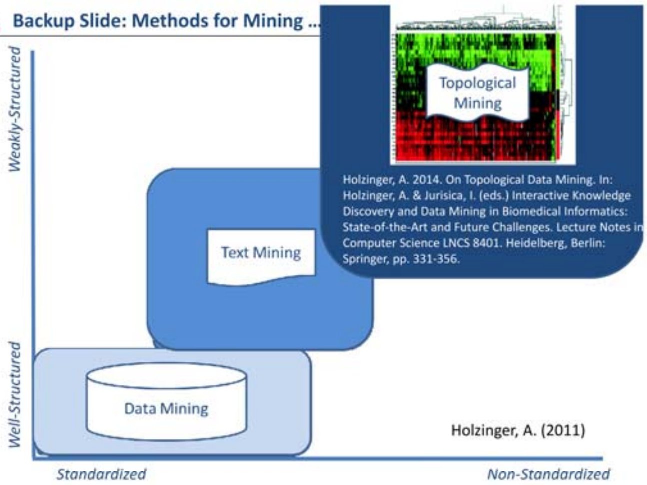
- **Hadoop Common**: The common utilities that support the other Hadoop subprojects.
- **Hadoop Distributed File System (HDFS™)**: A distributed file system that provides high-throughput access to application data.
- **Hadoop MapReduce**: A software framework for distributed processing of large data sets on compute clusters.

Other Hadoop-related projects at Apache include:

- **Avro™**: A data serialization system.
- **Cassandra™**: A scalable multi-master database with no single points of failure.
- **Chukwa™**: A data collection system for managing large distributed systems.
- **HBase™**: A scalable, distributed database that supports structured data storage for large tables.
- **Hive™**: A data warehouse infrastructure that provides data summarization and ad hoc querying.
- **Mahout™**: A Scalable machine learning and data mining library.
- **Pig™**: A high-level data-flow language and execution framework for parallel computation.
- **ZooKeeper™**: A high-performance coordination service for distributed applications.

Taylor, R. C. (2010) An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics. *BMC Bioinformatics, 11, 1-6.*

Backup Slide: Methods for Mining ...

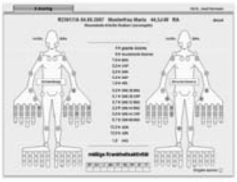


Holzinger, A. 2014. On Topological Data Mining. In: Holzinger, A. & Jurisica, I. (eds.) Interactive Knowledge Discovery and Data Mining in Biomedical Informatics: State-of-the-Art and Future Challenges. Lecture Notes in Computer Science LNCS 8401. Heidelberg, Berlin: Springer, pp. 331-356.

Holzinger, A. (2011)

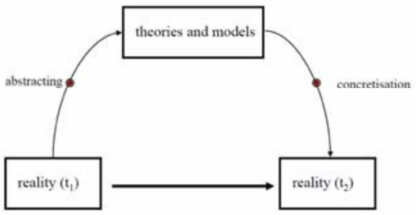
Slide 2-8: Example of typical clinical data sets

- 50+ Patients per day ~ 5000 data points per day …
- Aggregated with specific scores (Disease Activity Score, DAS)
- Current patient status is related to previous data
- = convolution over time
- ⇒ **time-series data**



Simonic, K. M., Holzinger, A., Bloice, M. & Hermann, J. (2011). *Optimizing Long-Term Treatment of Rheumatoid Arthritis with Systematic Documentation. Pervasive Health - 5th International Conference on Pervasive Computing Technologies for Healthcare, Dublin, IEEE, 550-554.*

Backup Slide: Excursion: How to get theories?



positivism :

$\{\text{theory, model}\} \notin \text{reality}$

$\text{reality}(t_1) \approx \text{reality}(t_2)$

constructionism :

$\{\text{theory, model}\} \in \text{reality}$

$\text{reality}(t_1) \neq \text{reality}(t_2)$

Rauterberg, M. (2006) HCI as an engineering discipline: to be or not to be. *African Journal of Information and Communication Technology, 2, 4, 163-184.*