Andreas Holzinger

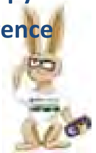
185.A83 Machine Learning for Health Informatics
2019S, VU, 2.0 h, 3.0 ECTS
Dienstag, 30. April 2019

From Data for Machine Learning to probabilistic information and entropy

andreas.holzinger@tuwien.ac.at
https://human-centered.ai/machine-learning-for-health-informatics-class-2019

Red thread through the lecture today

- 01 Data – the underlying physics of data
- 02 Biomedical data sources – taxonomy of data
- 03 Data integration, mapping, fusion
- 04 Probabilistic Information
- 05 Information Theory – Information Entropy
- 06 Cross- Entropy - Kullback-Leibler Divergence



01 Reflection

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

Warm-up Quiz

1

2

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

3

4

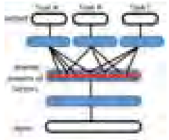
5

6

7

context

8



9

Question: Where is the Biologist in this image?

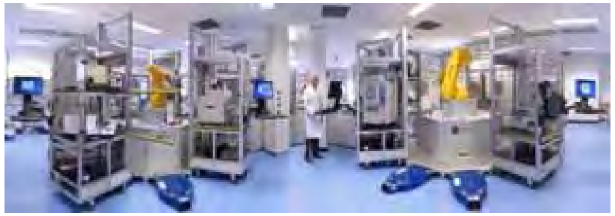


Image source: http://www.efmc.info/medchemwatch-2014-1/lab.php

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

Questions …

- What does “big data” mean? Is it good or bad?
- Would you collect more data of low quality?
- Or would you use only data of high quality?
- What is data quality?
- How do you measure data quality?
- What about data protection and privacy?
- What about data security?
- What does data accessibility mean?
- What does interpretability need?

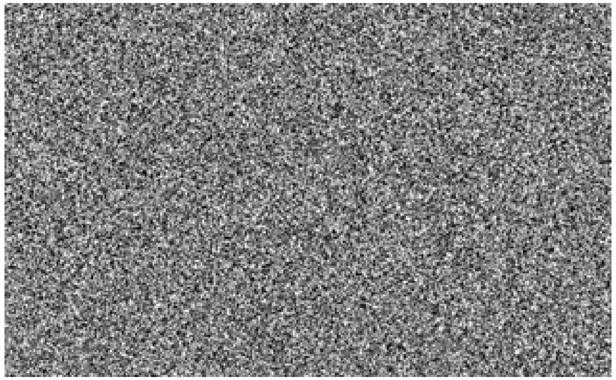
Leo L. Pipino, Yang W. Lee & Richard Y. Wang 2002. Data quality assessment. Communications of the ACM, 45, (4), 211-218.

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 in high-dimensions (manual handling of the data is 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.

What is this?



01 The underlying physics of data

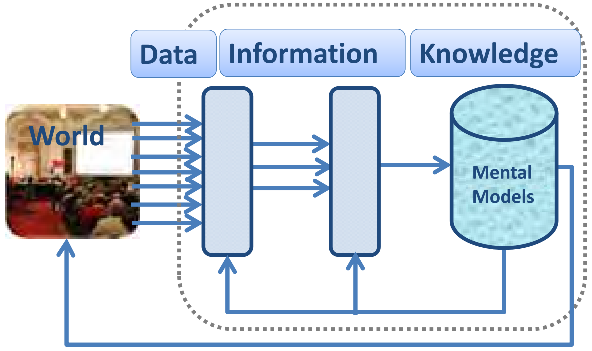
Traditional Statistics versus Machine Learning

- 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 can **not** be represented by a simple model

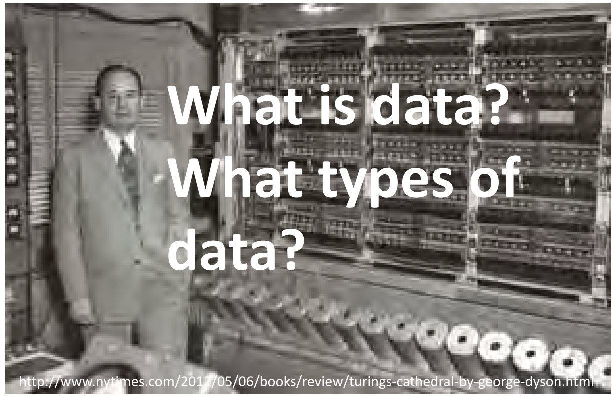
Lecun, Y., Bengio, Y. & Hinton, G. 2015. Deep learning. Nature, 521, (7553), 436-444.

Data – Information - Knowledge



Knowledge := a set of expectations

Institute for Advanced Study, Princeton University



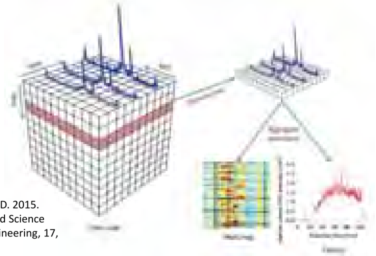
What is data? What types of data?

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

Example: Neonatal Screening (1/3)

http://www.nlm.nih.gov/cgi/mesh/2011/MB_cgi?mode=&index=15177&view=expanded#TreeE01.370.500.580

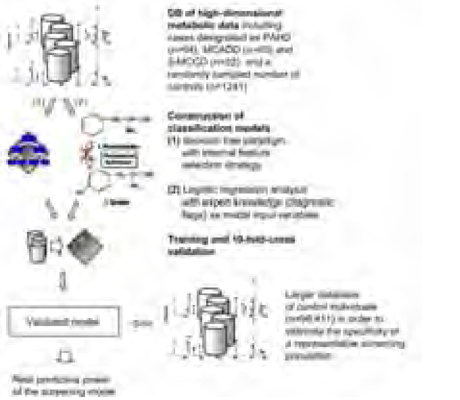
Example: Neonatal Screening (2/3)



Yao, Y., Bowen, B. P., Baron, D. & Poznanski, D. 2015. SciDB for High-Performance Array-Structured Science Data at NERSC. Computing in Science & Engineering, 17, (3), 44-52, doi:10.1109/MCSE.2015.43.

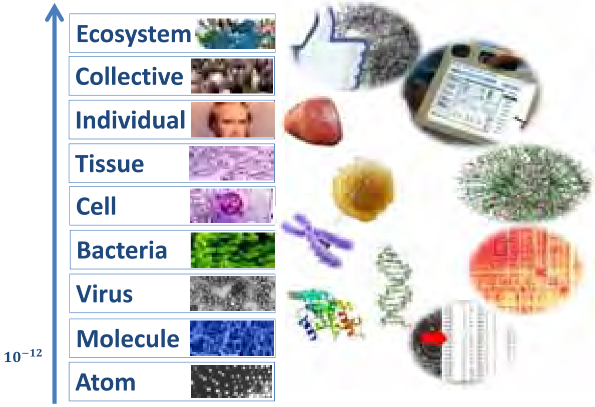
Example: Neonatal Screening (3/3)

Baumgartner, C., Bohm, C. & Baumgartner, D. 2005. Modelling of classification rules on metabolic patterns including machine learning and expert knowledge. Journal of Biomedical Informatics, 38, (2), 89-98, doi:10.1016/j.jbi.2004.08.009.

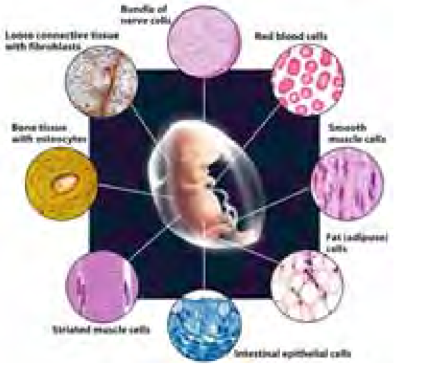

Construction of classification models
Training and 10-fold-cross validation
Validated model

02 Biomedical data sources: Taxonomy of data

Top Level View: Biomedical Data Sources



Example: Type of Cells

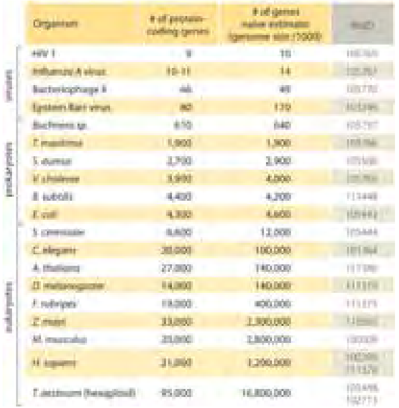


Karp, G. 2010. Cell and Molecular Biology: Concepts and Experiments, Gainesville, John Wiley.

To get a feeling of biological data sources (bionumbers)

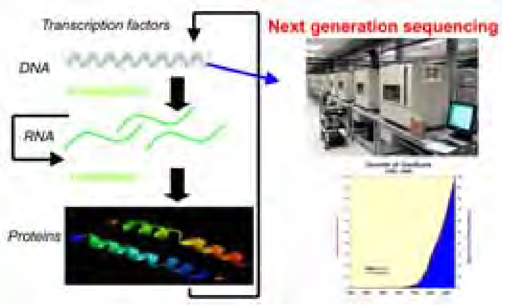
bionumbers.hms.harvard.edu/

http://book.bionumbers.
org/how-many-genes-
are-in-a-genome/



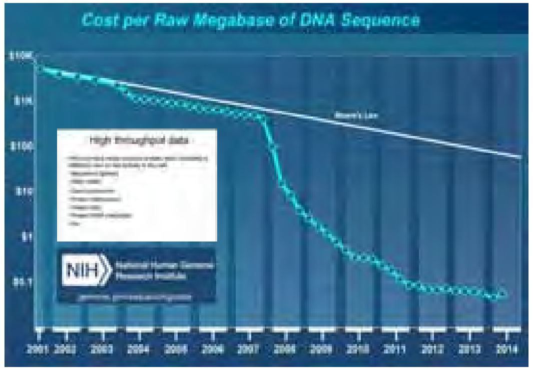
	Organism	# of protein-coding genes	# of genes naive estimate (genome size /1000)	BNID
viruses	HIV 1	9	10	[illegible]
	Influenza A virus	10-11	14	[illegible]
	Bacteriophage λ	66	49	[illegible]
	Epstein Barr virus	80	170	[illegible]
prokaryotes	Buchnera sp.	610	640	[illegible]
	T. maritima	1,900	1,900	[illegible]
	S. aureus	2,700	2,900	[illegible]
	V. cholerae	3,900	4,000	[illegible]
	B. subtilis	4,400	4,200	[illegible]
	E. coli	4,300	4,600	[illegible]
eukaryotes	S. cerevisiae	6,600	12,000	[illegible]
	C. elegans	20,000	100,000	[illegible]
	A. thaliana	27,000	140,000	[illegible]
	D. melanogaster	14,000	140,000	[illegible]
	F. rubripes	19,000	400,000	[illegible]
	Z. mays	33,000	2,300,000	[illegible]
	M. musculus	20,000	2,800,000	[illegible]
	H. sapiens	21,000	3,200,000	[illegible]
	T. aestivum (hexaploid)	95,000	16,800,000	[illegible]

Biological data is getting more complex (big sowieso ;)

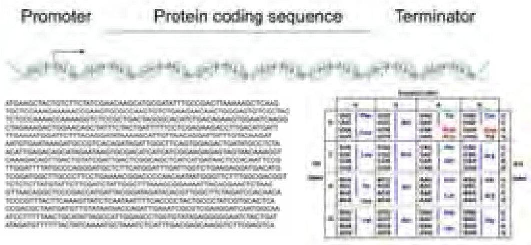


Navlakha, S. & Bar-Joseph, Z. 2011. Algorithms in nature: the convergence of systems biology and computational thinking. *Molecular Systems Biology,* 7.

Costs more decreasing than in Moore’s law (also cost!)

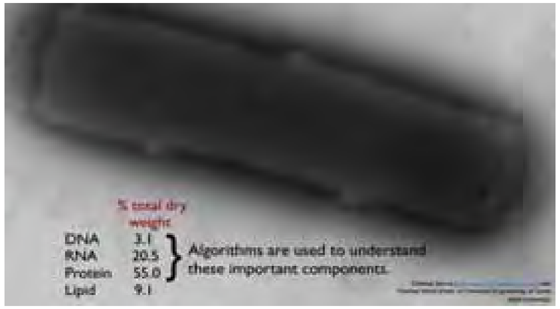


Example: Genetic Data



For further reading this is recommended:
Buffalo, V. 2015. Bioinformatics Data Skills: Reproducible and Robust Research with Open Source Tools, Sebastopol (CA), O'Reilly.

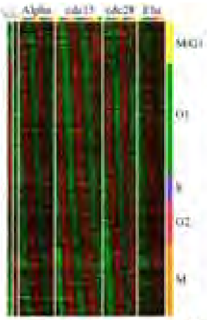
Example Species: Bacterium E. coli


% total dry weight
DNA 3.1
RNA 20.5
Protein 55.0
Lipid 9.1
Algorithms are used to understand these important components.

Where do we get the data sets from?

- Billions of biological data sets are openly available, here only some examples:
- General Repositories:
  - GenBank, EMBL, HMCA, …
- Specialized by data types:
  - UniProt/SwissProt, MMMP, KEGG, PDB, …
- Specialized by organism:
  - WormBase, FlyBase, NeuroMorpho, …
- Details: http://hci-kdd.org/open-data-sets

Example Data Set from an High-Throughput Experiment



- this figure depicts one yeast gene-expression data set
  - each row represents a gene
  - each column represents a measurement of gene expression (mRNA abundance) at some time point
    - red indicates that a gene is being expressed more than some baseline; green means less

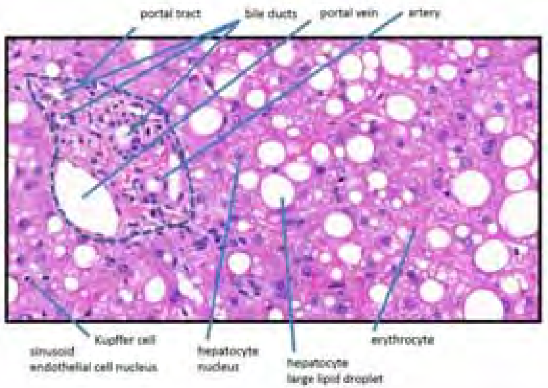
Figure from Spellman et al., Molecular Biology of the Cell, 9:3273-3297, 1998

Taxonomy of data

- **Physical level** -> bit = binary digit = **b**asic **i**ndissoluble 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, numerical measurements (physiological data, lab results, vital signs, …), recorded signals (ECG, EEG, …), Images (x-ray, MR, CT, PET, …) ; -omics

Taxonomy of data at Hospital Level

- **Clinical workplace data sources**
  - Medical documents: text (non-standardized (“free-text”), semi-structured, standard terminologies (ICD, SNOMED-CT)
  - Measurements: lab, time series, ECG, EEG, EOG, …
  - Surveys, Clinical study data, trial data
- **Image data sources**
  - Radiology: MRI (256x256, 200 slices, 16 bit per pixel, uncompressed, ~26 MB); CT (512x512, 60 slices, 16 bit per pixel, uncompressed ~32MB; MR, US;
  - Digital Microscopy : WSI (15mm slide, 20x magn., 24 bits per pixel, uncompressed, 2,5 GB, WSI 10 GB; confocal laser scanning, etc.
- **-omics data sources**
  - Sanger sequencing, NGS whole genome sequencing (3 billion reads, read length of 36) ~ 200 GB; NGS exome sequencing (“only” 110,000,000 reads, read length of 75) ~7GB; Microarray, mass-spectrometry, gas chromatography, …



Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal & Heimo Mueller 2019. Causability and Explainability of AI in Medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, doi:10.1002/widm.1312.

Causability must ensure the quality of explanations

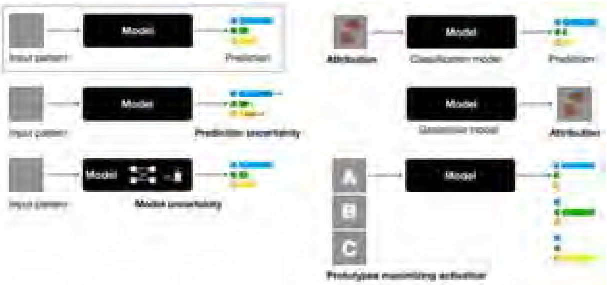
Level 1 Association $P(y|x)$ with the typical activity of "seeing" and questions including "How would seeing X change my belief in Y?", in our use-case above this was the question of "what does a feature in a histology slide the pathologist about a disease?"

Level 2 Intervention $P(y|do(x), z)$ with the typical activity of "doing" and questions including "What if I do X?", in our use-case above this was the question of "what if the medical professional recommends treatment X - will the patient be cured?"

Level 3 Counterfactuals $P(y_x|x', y')$ with the typical activity of "retrospection" and questions including "Was Y the cause for X?", in our use-case above this was the question of "was it the treatment that cured the patient?"

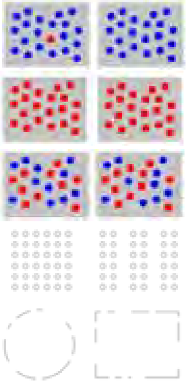
Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal & Heimo Mueller 2019. Causability and Explainability of AI in Medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, doi:10.1002/widm.1312.

Our solution is in weakly-supervised learning

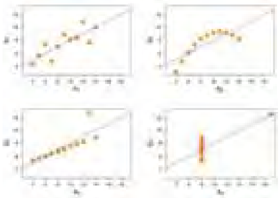


Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal & Heimo Mueller 2019. Causability and Explainability of AI in Medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, doi:10.1002/widm.1312.

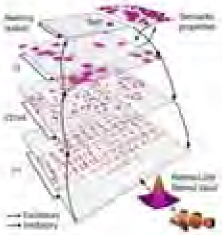
Gestalt Laws: Proximity, Similarity, Closure, Symmetry, …

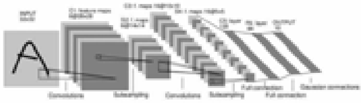


Kurt Koffka 1935. Principles of Gestalt Psychology, New York, Harcourt.



https://en.wikipedia.org/wiki/Anscombes_quartet

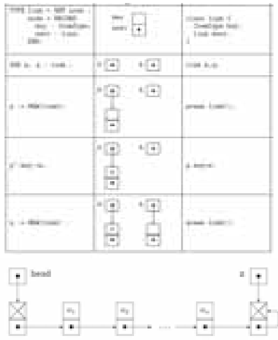


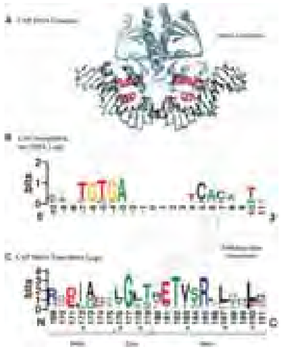


David H. Hubel & Torsten N. Wiesel 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. The Journal of Physiology, 160, (1), 106-154, doi:10.1113/jphysiol.1962.sp006837.

Yann Lecun, Bernhard Boser, John S. Denker, Donnie Henderson, Richard E. Howard, Wayne Hubbard & Lawrence D. Jackel 1989. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1, (4), 541-551, doi:10.1162/neco.1989.1.4.541.

Example Data Structures (1/3): List

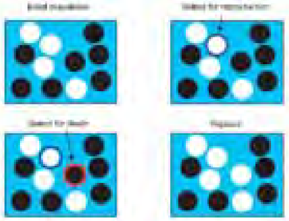


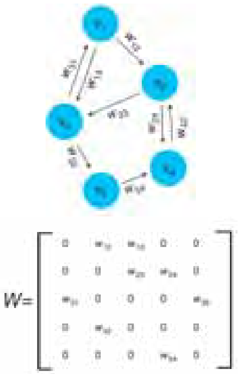


Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. (2004) WebLogo: A sequence logo generator. Genome Research, 14, 6, 1188-1190.

Example Data Structures (2/3): Graph

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

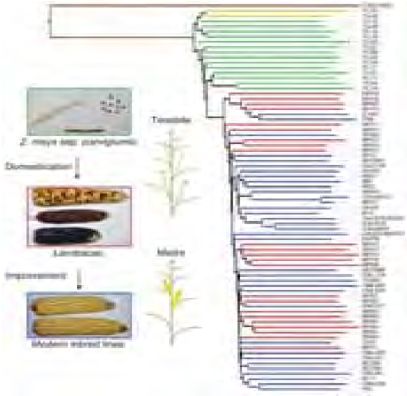




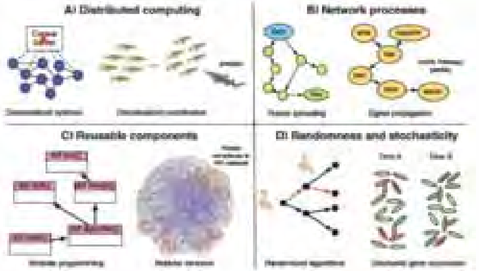
Lieberman, E., Hauert, C. & Nowak, M. A. (2005) Evolutionary dynamics on graphs. *Nature, 433, 7023, 312-316.*

Example Data Structures (3/3) Tree

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



Algorithms in nature: Shared principles



http://cacm.acm.org/magazines/2015/1/181614-distributed-information-processing-in-biological-and-computational-systems/abstract

Navlakha, S. & Bar-Joseph, Z. 2014. Distributed information processing in biological and computational systems. Commun. ACM, 58, (1), 94-102.

https://www.youtube.com/watch?v=4u47nwHzqI4&feature=youtu.be

Biomedical Research needs Open Data Access

- Grand Challenges in this area:
- – Production of Open Data Sets
- – Synthetic data sets for learning algorithm testing
- – Privacy preserving machine learning
- – Data leak detection
- – Data citation
- – Differential privacy
- – Anonymization and pseudonymization
- – Evaluation and benchmarking

Please visit:
http://hci-kdd.org/privacy-aware-machine-learning-for-data-science/

03 Data Integration, mapping, fusion

Unsolved Problem: Data Integration and Data Fusion in the Life Sciences

How to combine these different data types together to obtain a unified view of the activity in the cell is one of the major challenges of systems biology

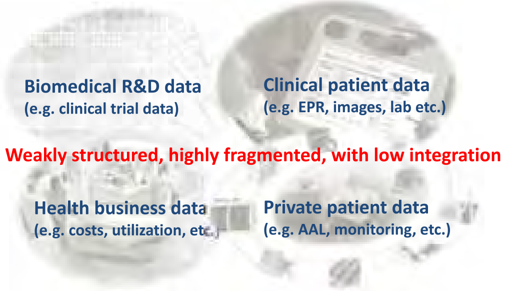
Navlakha, S. & Bar-Joseph, Z. 2014. Distributed information processing in biological and computational systems. *Commun. ACM*, 58, (1), 94-102, doi:10.1145/2678280.

In medicine we have two different worlds …

Our central hypothesis: Information may bridge this gap

Holzinger, A. & Simonic, K.-M. (eds.) 2011. *Information Quality in e-Health. Lecture Notes in Computer Science LNCS 7058, Heidelberg, Berlin, New York: Springer.*

Not only a problem at cell level …



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

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

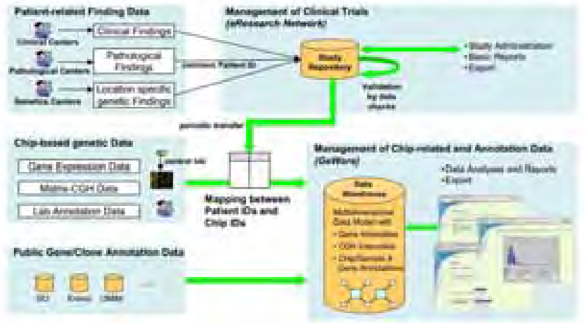
Weakly structured, highly fragmented, with low integration

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

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

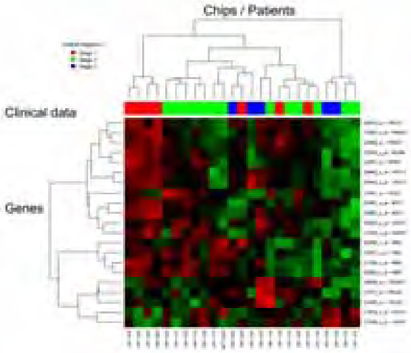
Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. & Byers, A. H. (2011) *Big data: The next frontier for innovation, competition, and productivity. Washington (DC), McKinsey Global Institute.*

Example Data Integration Architecture


Patient-related Finding Data
Clinical Centers
Clinical Findings
Pathological Centers
Pathological Findings
Genetics Centers
Location specific genetic Findings
Management of Clinical Trials (eResearch Network)
Study Repository
•Study Administration
•Basic Reports
•Export
periodic transfer
Chip-based genetic Data
Gene Expression Data
Matrix CGH Data
Lab Annotation Data
Mapping between Patient IDs and Chip IDs
Management of Chip-related and Annotation Data (GeWare)
•Data Analyses and Reports
•Export
Data Warehouse
Public Gene/Clone Annotation Data

Kirsten, T., Lange, J. & Rahm, E. 2006. An integrated platform for analyzing molecular-biological data within clinical studies. Current Trends in Database Technology–EDBT 2006. Heidelberg: Springer, pp. 399-410, doi:10.1007/11896548_31.

Example: Integrated Data Set

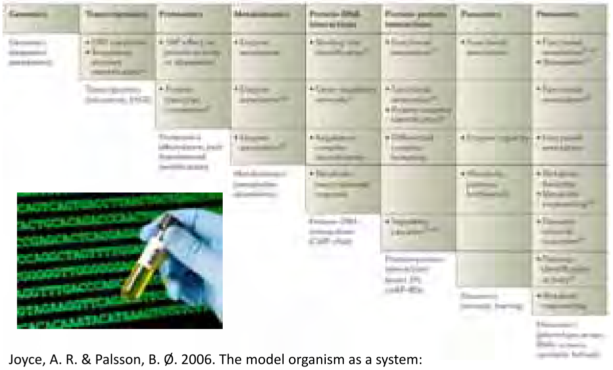


Kirsten, T., Lange, J. & Rahm, E. 2006. An integrated platform for analyzing molecular-biological data within clinical studies. Current Trends in Database Technology–EDBT 2006. Heidelberg: Springer, pp. 399-410, doi:10.1007/11896548_31.

-Omics-data jungle

- **Gen**omics (sequence annotation)
- **Transcript**omics (microarray)
- **Prote**omics (Proteome Databases)
- **Metabol**omics (enzyme annotation)
- **Flux**omics (isotopic tracing, metabolic pathways)
- **Phen**omics (biomarkers)
- **Epigen**omics (epigenetic modifications)
- **Microbi**omics (microorganisms)
- **Lipid**omics (pathways of cellular lipids)

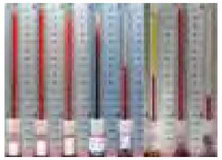
Omics-data integration

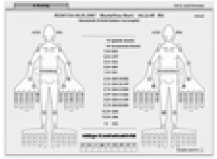


Joyce, A. R. & Palsson, B. Ø. 2006. The model organism as a system: integrating'omics' data sets. *Nature Reviews Molecular Cell Biology, 7***, 198-210.**

Example of typical clinical data sets

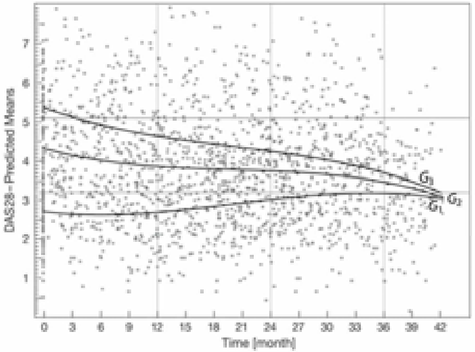
- 50+ Patients per day ~ 5000 data points per day …
- Aggregated with specific scores (Disease Activity Score, DAS)
- Current patient status is related to previous data
- = convolution over time
- ⇒ **time-series data**





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.*

DAS28 Predicted Mean Responses



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.*

Data Dimensionality examples

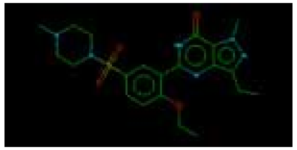
- 0-D data = a data point existing isolated from other data, e.g. integers, letters, Booleans, etc.
- 1-D data = consist of a string of 0-D data, e.g. Sequences representing nucleotide bases and amino acids, SMILES etc.
- 2-D data = having spatial component, such as images, NMR-spectra etc.
- 2.5-D data = 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:

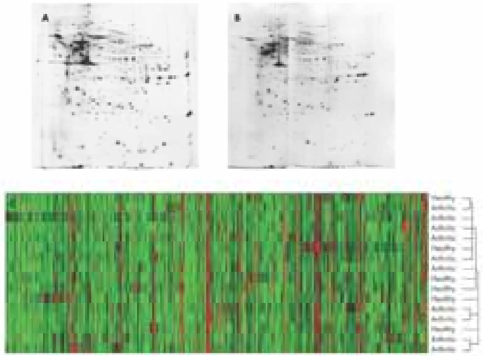
e.g. Viagra:
CCc1nn(C)c2c(=O)[nH]c(nc12)c3cc(ccc3OCC)S(=O)(=O)N4CCN(C)CC4



…is Canonicalizable
…is Comprehensive
…is Well Documented

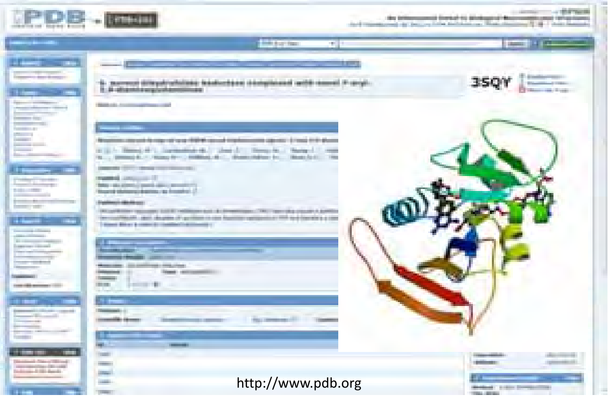
http://www.daylight.com/dayhtml_tutorials/languages/smiles/index.html

Example: 2-D data (bivariate data)



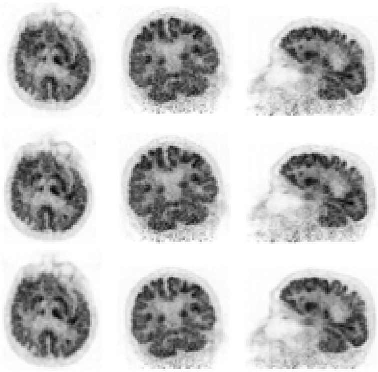
Kastrinaki et al. (2008) Functional, molecular & proteomic characterisation of bone marrow mesenchymal stem cells in rheumatoid arthritis. *Annals of Rheumatic Diseases, 67, 6, 741-749.*

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



http://www.pdb.org

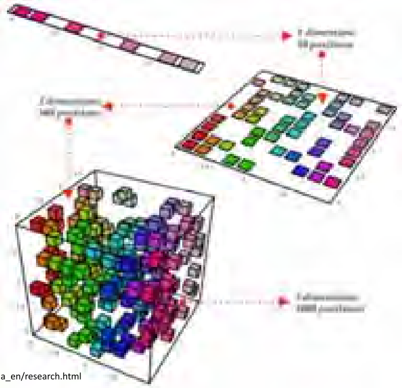
Example: 3-D Voxel data (volumetric picture elements)



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.*

Note: The curse of dimensionality

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.


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

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

Challenges in Data Integration

- Bridging the gap between natural sciences and clinical medicine (who has seen genomics and patient data integrated in routine???)
- Organizational barriers, data provenance, data ownership, privacy, accessibility, usability, fair use of data, security, safety, data protection
- Combine Ontologies with Machine Learning
- Stochastic Ontologies, Ontology learning
- Integration of data from wet-labs with in-silico experimental data (e.g. tumor growth simulation)

04 Probabilistic Information p(x)

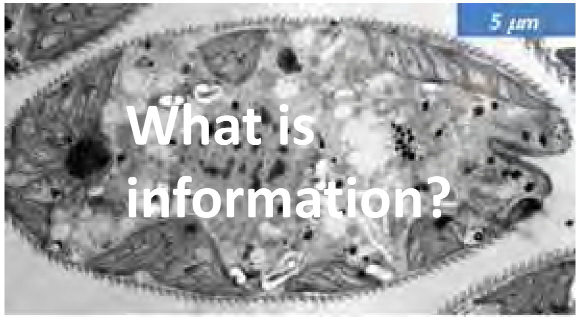
Please always distinguish models between …

- **Boolean models**
- **Algebraic models**
- **Probabilistic models *)**

*) Our probabilistic models describes data which we can observe from our environment – and if we use the mathematics of probability theory , in order to express the uncertainties around our model then the inverse probability allows us to infer unknown unknowns … learning from data and making predictions – the core essence of machine learning and of vital importance for health informatics

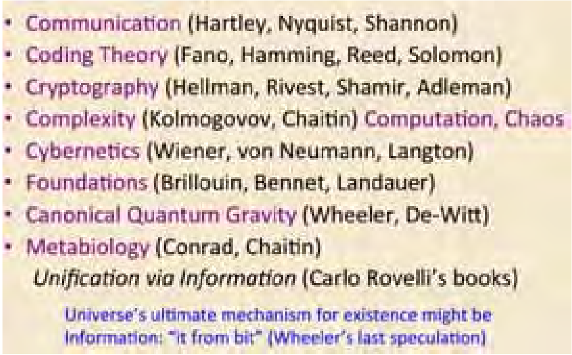
Ghahramani, Z. 2015. Probabilistic machine learning and artificial intelligence. Nature, 521, (7553), 452-459, doi:10.1038/nature14541.

Life is complex information



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

Information is everywhere …



- Communication (Hartley, Nyquist, Shannon)
- Coding Theory (Fano, Hamming, Reed, Solomon)
- Cryptography (Hellman, Rivest, Shamir, Adleman)
- Complexity (Kolmogovov, Chaitin) Computation, Chaos
- Cybernetics (Wiener, von Neumann, Langton)
- Foundations (Brillouin, Bennet, Landauer)
- Canonical Quantum Gravity (Wheeler, De-Witt)
- Metabiology (Conrad, Chaitin)

Unification via Information (Carlo Rovelli's books)

Universe's ultimate mechanism for existence might be information: "It from bit" (Wheeler's last speculation)

Manca, V. 2013. Infobiotics: Information in Biotic Systems, Heidelberg, Springer, doi:10.1007/978-3-642-36223-1.

For ML and Health always remember

Probabilistic Information p(x)

Bayes, T. (1763). An Essay towards solving a Problem in the Doctrine of Chances (Postum communicated by Richard Price). Philosophical Transactions, 53, 370-418.

$$p(x_i) = \sum P(x_i, y_j)$$

Thomas Bayes
1701 - 1761

$$p(x_i, y_j) = p(y_j|x_i)P(x_i)$$

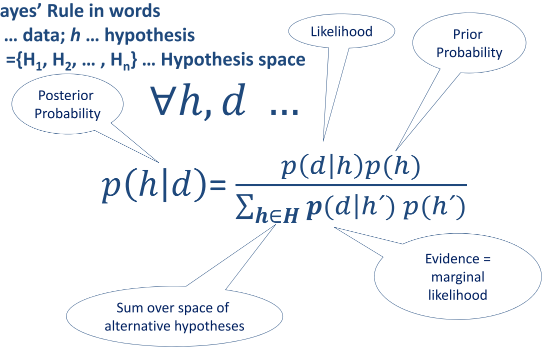
Bayes' Rule is a corollary of the Sum Rule and Product Rule:

$$p(x_i|y_j) = \frac{p(y_j|x_i)p(x_i)}{\sum p(x_i, y_j)p(x_i)}$$

Barnard, G. A., & Bayes, T. (1958). Studies in the history of probability and statistics: IX. Thomas Bayes's essay towards solving a problem in the doctrine of chances. Biometrika, 45(3/4), 293-315.

Bayes Law of Total Probability = data modelling

Bayes' Rule in words
d **… data;** ***h*** **… hypothesis**
H ={H$_1$, H$_2$, … , H$_n$} … Hypothesis space



$$\forall h, d \;\; \dots$$

$$p(h|d) = \frac{p(d|h)p(h)}{\sum_{h \in H} p(d|h') \, p(h')}$$

Likelihood
Prior Probability
Posterior Probability
Evidence = marginal likelihood
Sum over space of alternative hypotheses

Always remember:

The inverse probability allows to infer unknowns, learn from data and make **predictions:**

1) Maximum-Likelihood Learning

finds a parameter setting, that maximizes the p(x) of the data: $P\,(\mathcal{D}\,|\theta\,)$

2) Maximum a Posteriori Learning (e.g. for MCMC)

assumes a prior over the model parameters $P\,(\theta)$ and finds a parameter setting that maximizes the posterior: $P(\theta\,|\mathcal{D}) \;\propto P(\theta)P(\mathcal{D}\,|\theta\,)$

3) Bayesian Learning

assumes a prior over the model parameters and computes the posterior distribution $P(\theta\,|\mathcal{D})$

Parameter Estimation

- General setting:
  - Given a (hypothesized & probabilistic) model that governs the random experiment
  - The model gives a probability of any data $p(D|\text{⍰})$ that depends on the parameter θ
  - Now, given actual sample data $X = \{x_1, \dots, x_n\}$, what can we say about the value of θ?
- Intuitively, take your best guess of θ
- "best" means "best explaining/fitting the data"
- Generally an optimization problem

Maximum Likelihood vs. Bayesian

- 1) Maximum likelihood estimation (given X)
  - "Best" means "data likelihood reaches maximum"

$$\hat{\theta} = \underset{\theta}{\arg\max}\, P(X|\theta)$$

  - **Problem: massive amount of data necessary**
- 2) Bayesian estimation (use posterior)

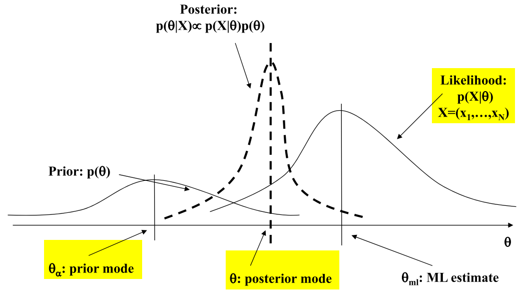
$$\hat{\theta} = \underset{\theta}{\arg\max}\, P(X|\theta) = \underset{\theta}{\arg\max}\, P(X|\theta)\, P(\theta)$$

  - "Best" means being consistent with our "prior" knowledge and explaining data well
  - **Problem: how to define prior?**

An example can be found in: Banerjee, O., El Ghaoui, L. & D'aspremont, A. 2008. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *The Journal of Machine Learning Research,* 9, 485-516. Available via: http://arxiv.org/pdf/0707.0704

Illustration of Bayesian Estimation

$$posterior\ p(x) = \frac{likelihood * prior\ p(x)}{evidence}$$



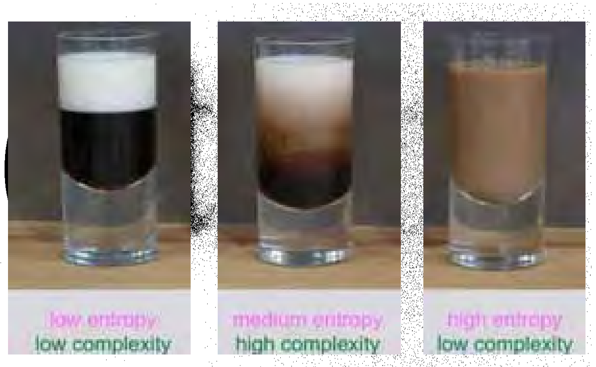
For more basic information: Bishop, C. M. 2007. *Pattern Recognition and Machine Learning,* Springer. For application examples in Text processing refer to: Jiang, J. & Zhai, C. X. 2007. An empirical study of tokenization strategies for biomedical information retrieval. *Information Retrieval,* 10, (4-5), 341-363.

05 Information Theory & Entropy

Probability > Information > Entropy

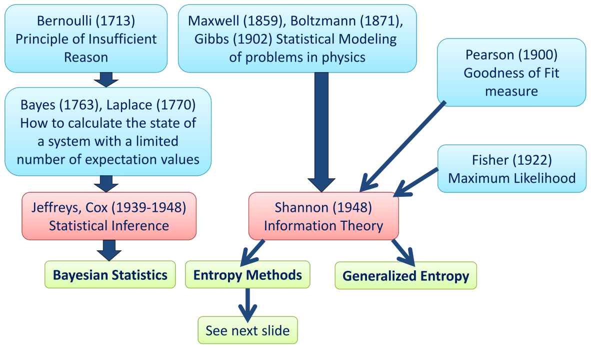
- Information is the reduction of uncertainty
- If something is 100 % certain its uncertainty = 0
- Uncertainty is max. if all choices are equally probable (I.I.D)
- Uncertainty (as information) sums up for independent sources

Entropy as measure for disorder



http://www.scottaaronson.com

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

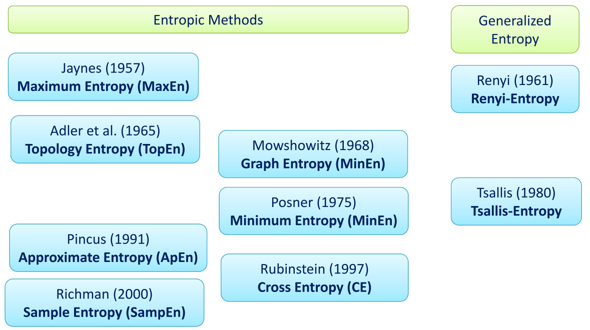
Entropy Methods

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)**

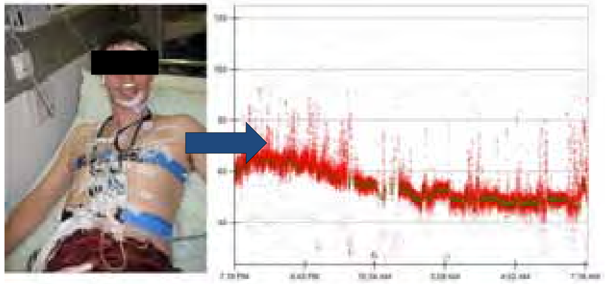
Pincus (1991) **Approximate Entropy (ApEn)**

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)

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

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

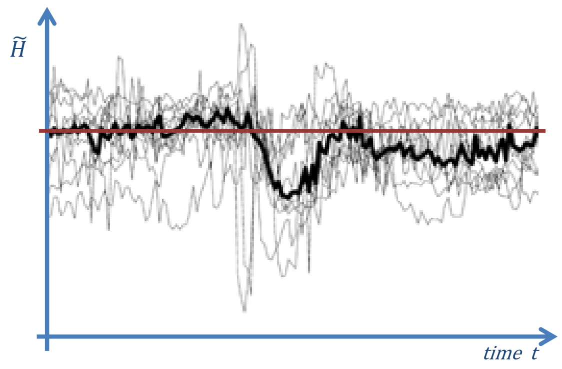
$$\left\| \vec{X}_i , \vec{X}_j \right\| = \max_{k=1,2,\dots,m} \left(\left| x_{(i+k-1)} - x_{(j+k-1)} \right| \right)$$

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

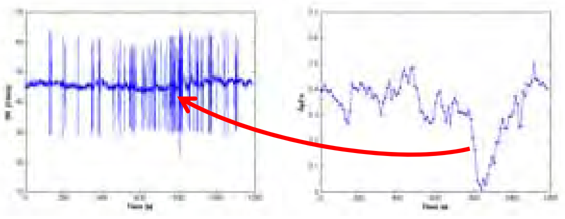
$$C_r^m(i) = \frac{N^m(i)}{N-m+1} \qquad \phi^m(r) = \frac{1}{N-m+1} \sum_{t=1}^{N-m+1} \ln C_r^m(i)$$

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.*

Example: ApEn (2)

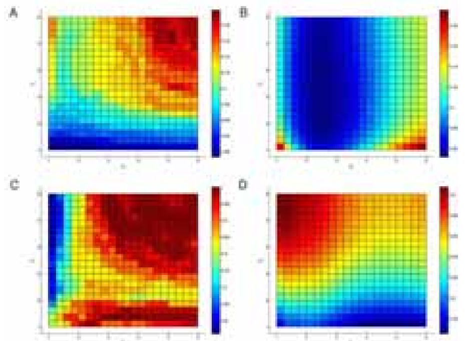


ApEn



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.

Significance of FuzzyMEn for different nL and nF, N=1000



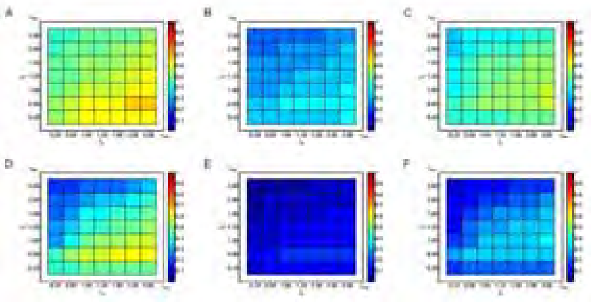
Mayer, C., Bachler, M., Hortenhuber, M., Stocker, C., Holzinger, A. & Wassertheurer, S. 2014. Selection of entropy-measure parameters for knowledge discovery in heart rate variability data. *BMC Bioinformatics,* 15, (Suppl 6), S2, doi:doi:10.1186/1471-2105-15-S6-S2.

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



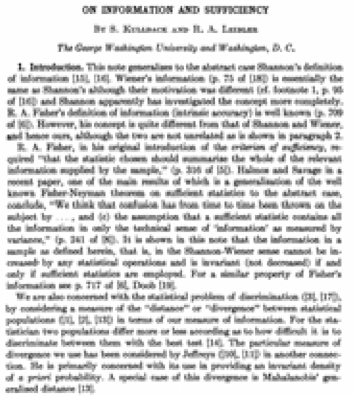
Mayer, C., Bachler, M., Holzinger, A., Stein, P. K. & Wassertheurer, S. 2016. The Effect of Threshold Values and Weighting Factors on the Association between Entropy Measures and Mortality after Myocardial Infarction in the Cardiac Arrhythmia Suppression Trial (CAST). Entropy, 18, (4), 129, doi::10.3390/e18040129.

06 Cross-Entropy Kullback-Leibler Divergence

Entropy – KL-Div. – Mutual Information

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

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].

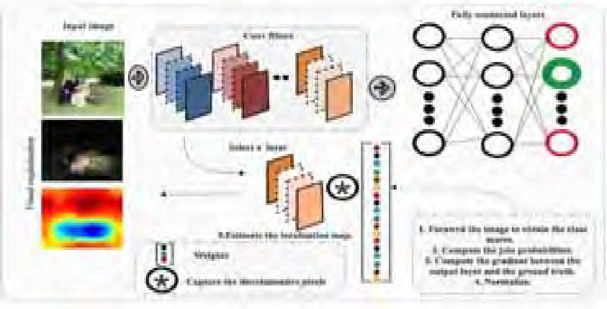
We are also concerned with the statistical problem of discrimination ([3], [17]), by considering a measure of the "distance" or "divergence" between statistical populations ([1], [2], [13]) in terms of our measure of information. For the statistician two populations differ more or less according as to 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

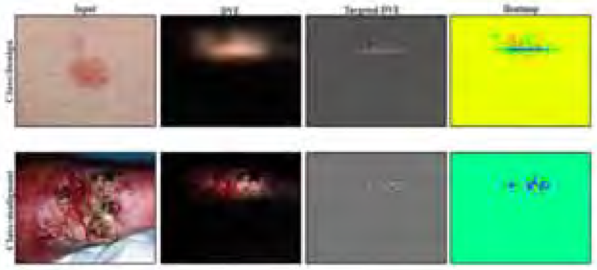
Richard Leibler
1914-2003

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

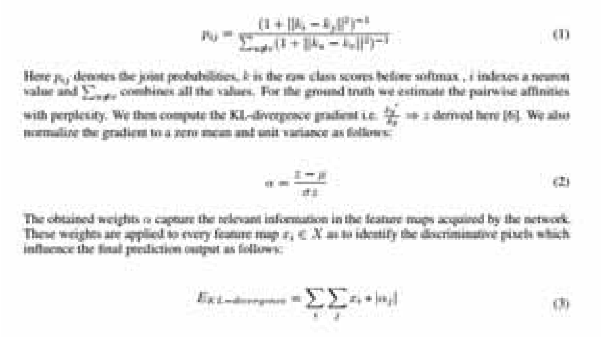
Deep Learning Explainability


Input image
Conv Blocks
Fully connected layers
Select a layer
3.Estimate the localization map
Weights
Capture the discriminative pixels
1. Forward the image to obtain the class scores.
2. Compute the join probabilities.
3. Compute the gradient between the output layer and the ground truth.
4. Normalize.

Housam Khalifa Bashier Babiker & Randy Goebel 2017. Using KL-divergence to focus Deep Visual Explanation. arXiv preprint arXiv:1711.06431.



Housam Khalifa Bashier Babiker & Randy Goebel 2017. An Introduction to Deep Visual Explanation. arXiv preprint arXiv:1711.09482.



$$p_{ij} = \frac{(1+\|k_i - k_j\|^2)^{-1}}{\sum_{n\neq m}(1+\|k_n - k_m\|^2)^{-1}} \quad (1)$$

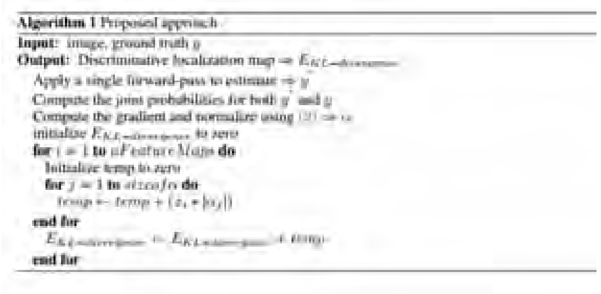
Here p_{ij} denotes the joint probabilities, k is the raw class scores before softmax , i indexes a neuron value and $\sum_{n\neq m}$ combines all the values. For the ground truth we estimate the pairwise affinities with perplexity. We then compute the KL-divergence gradient i.e. $\frac{\delta y'}{\delta y} \Rightarrow z$ derived here [6]. We also normalize the gradient to a zero mean and unit variance as follows:

$$\alpha = \frac{z-\mu}{\sigma z} \quad (2)$$

The obtained weights α capture the relevant information in the feature maps acquired by the network. These weights are applied to every feature map $x_i \in X$ as to identify the discriminative pixels which influence the final prediction output as follows:

$$E_{KL-divergence} = \sum_i \sum_j x_i * |\alpha_j| \quad (3)$$

Housam Khalifa Bashier Babiker & Randy Goebel 2017. Using KL-divergence to focus Deep Visual Explanation. arXiv preprint arXiv:1711.06431.



```
Algorithm 1 Proposed approach
Input: image, ground truth $y$
Output: Discriminative localization map $\Rightarrow E_{KL-divergence}$
  Apply a single forward-pass to estimate $\Rightarrow y'$
  Compute the joint probabilities for both $y'$ and $y$
  Compute the gradient and normalize using (2) $\Rightarrow \alpha$
  initialize $E_{KL-divergence}$ to zero
  for $i = 1$ to $\#FeatureMaps$ do
    Initialize temp to zero
    for $j = 1$ to $\#\alpha$ do
      $temp \leftarrow temp + (x_i * |\alpha_j|)$
  end for
    $E_{KL-divergence} \leftarrow E_{KL-divergence} + temp$
  end for
```

Housam Khalifa Bashier Babiker & Randy Goebel 2017. Using KL-divergence to focus Deep Visual Explanation. arXiv preprint arXiv:1711.06431.

Remember Shannon Entropy

$$\mathrm{H}[x] = -\sum_x p(x)\log_2 p(x)$$

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

Important quantity in
- coding theory
- statistical physics
- machine learning

Conditional Entropy

$$\mathrm{H}[\mathbf{y}|\mathbf{x}] = -\iint p(\mathbf{y},\mathbf{x}) \ln p(\mathbf{y}|\mathbf{x})\,\mathrm{d}\mathbf{y}\,\mathrm{d}\mathbf{x}$$

$$\mathrm{H}[\mathbf{x},\mathbf{y}] = \mathrm{H}[\mathbf{y}|\mathbf{x}] + \mathrm{H}[\mathbf{x}]$$

The Kullback-Leibler Divergence

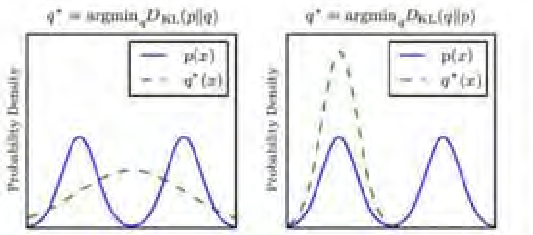
$$\begin{aligned}\mathrm{KL}(p\|q) &= -\int p(\mathbf{x})\ln q(\mathbf{x})\,\mathrm{d}\mathbf{x} - \left(-\int p(\mathbf{x})\ln p(\mathbf{x})\,\mathrm{d}\mathbf{x}\right) \\ &= -\int p(\mathbf{x})\ln\left\{\frac{q(\mathbf{x})}{p(\mathbf{x})}\right\}\mathrm{d}\mathbf{x}\end{aligned}$$

$$\mathrm{KL}(p\|q) \simeq \frac{1}{N}\sum_{n=1}^{N}\{-\ln q(\mathbf{x}_n|\boldsymbol{\theta}) + \ln p(\mathbf{x}_n)\}$$

$$\mathrm{KL}(p\|q) \geqslant 0$$

KL-divergence is often used to measure the distance between two distributions

Note: KL is not symmetric!



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

Goodfellow, I., Bengio, Y. & Courville, A. 2016. Deep Learning, Cambridge (MA), MIT Press.

Entropy measures generally …

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

Thank you!

Questions

Sample Questions (1)

- What are the grand challenges in ML for health?
- What is the key problem before you can apply ML?
- Describe the taxonomy of data at Hospital level!
- What does translational medicine mean?
- Give an example for a 2.5D-data set!
- Why would be the combination of ontologies with machine learning provide a benefit?
- How did Van Bemmel and Musen describe the interplay between data-information-knowledge?
- What is the "body-of-knowledge" in medical jargon?
- How do human process information?

Sample Questions (2)

- What was our definition of "knowledge"?
- What is the huge benefit of a probabilistic model?
- Please explain Bayes law with view on ML!
- What is information in the sense of Shannon?
- Why is information theory for us important?
- Which benefits provide entropic methods for us?
- Why is feature selection so important?
- What can you do with the Kullback-Leibler Divergence?

Appendix

Mutual Information

and Point Wise MI

Mutual Information I(X;Y): Comparing two distributions

$$\mathrm{I}[\mathbf{x},\mathbf{y}] \equiv \mathrm{KL}(p(\mathbf{x},\mathbf{y})\|p(\mathbf{x})p(\mathbf{y})) = -\iint p(\mathbf{x},\mathbf{y}) \ln\left(\frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x},\mathbf{y})}\right) \mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y}$$

$$\mathrm{I}[\mathbf{x},\mathbf{y}] = \mathrm{H}[\mathbf{x}] - \mathrm{H}[\mathbf{x}|\mathbf{y}] = \mathrm{H}[\mathbf{y}] - \mathrm{H}[\mathbf{y}|\mathbf{x}]$$

- Measures how much reduction in uncertainty of X given the information about Y
- Measures correlation between X and Y
- Related to the "channel capacity" in the original Shannon information theory

Bishop, C. M. 2007. *Pattern Recognition and Machine Learning*, Heidelberg, Springer.

Example: Disease-Disease Relationship

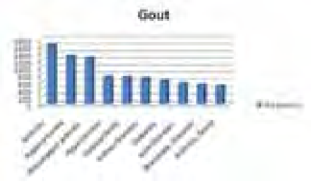
Let two words, w_i and w_j, have probabilities $P(w_i)$ and $P(w_j)$. Then their mutual information *PMI* (w_j, w_i) is defined as:

$$PMI(w_i, w_j) = \log\left(\frac{P(w_i, w_j)}{P(w_i)\,P(w_j)}\right)$$

For w_i denoting *rheumatoid arthritis* and w_j representing *diffuse scleritis* the following simple calculation yields:

$$P(w_i) = \frac{94{,}834}{[illegible]} \quad P(w_j) = \frac{74}{[illegible]}$$

$$P(w_i, w_j) = \frac{11}{94{,}834} \quad PMI(w_i, w_j) = 7.7.$$



Holzinger, A., Simonic, K. M. & Yildirim, P. Disease-Disease Relationships for Rheumatic Diseases: Web-Based Biomedical Textmining an Knowledge Discovery to Assist Medical Decision Making. 36th Annual IEEE Computer Software and Applications Conference (COMPSAC), 16-20 July 2012 2012 Izmir. IEEE, 573-580, doi:10.1109/COMPSAC.2012.77.

$$SCP(x,y) = p(x|y) \cdot p(y|x) = \frac{p(x,y)}{p(y)} \cdot \frac{p(x,y)}{p(x)} = \frac{p(x,y)^2}{p(x) \cdot p(y)}$$

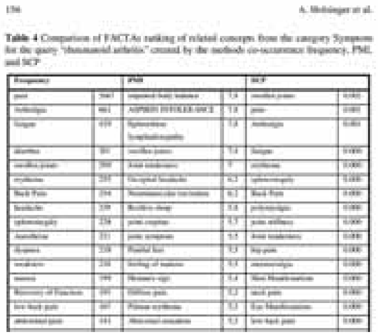

156 A. Holzinger et al.

Table 4 Comparison of FACTAs ranking of related concepts from the category Symptoms for the query "rheumatoid arthritis" created by the methods co-occurrence frequency, PMI, and SCP

Frequency		PMI		SCP	
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Back Pain	[illegible]	[illegible]	[illegible]	Back Pain	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Holzinger, A., Yildirim, P., Geier, M. & Simonic, K.-M. 2013. Quality-Based Knowledge Discovery from Medical Text on the Web. In: Pasi, G., Bordogna, G. & Jain, L. C. (eds.) Quality Issues in the Management of Web Information, Intelligent Systems Reference Library, ISRL 50. Berlin Heidelberg: Springer, pp. 145-158, doi:10.1007/978-3-642-37688-7_7.

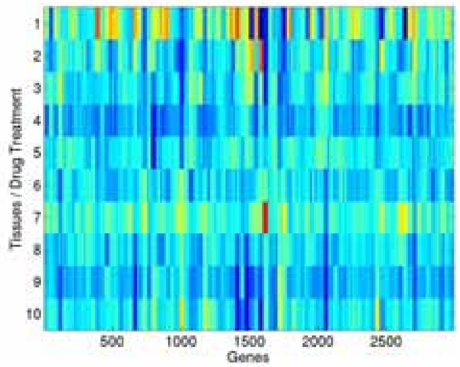
Open Questions, future outlook, hot topics, challenges

- 1) Challenges include –omics data analysis, where KL divergence and related concepts could provide important **measures** for discovering biomarkers.
- 2) Hot topics are new entropy measures suitable for computations in the context of complex/uncertain data for ML algorithms.
- Inspiring is the abstract geometrical setting underlying ML main problems, e.g. Kernel functions can be completely understood in this perspective. Future work may include entropic concepts and geometrical settings.

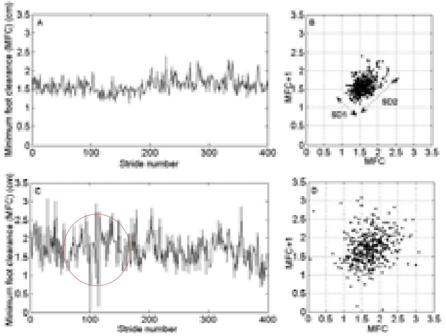
Limitations and Open Problems

- The case of higher order statistical structure in the data – nonlinear and hierarchical ?
- Outliers in the data – noise models?
- There are $\frac{D(D+1)}{2}$ parameters in a multi-variate Gaussian model – what happens if $D \gg$? dimensionality reduction

Thank you!

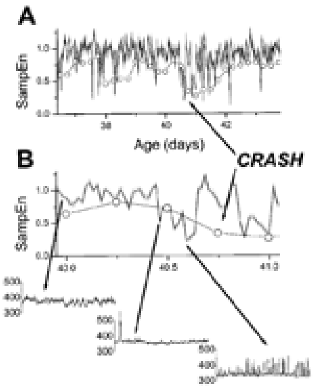


Back-up Slide: Poincare Plot for gait analysis



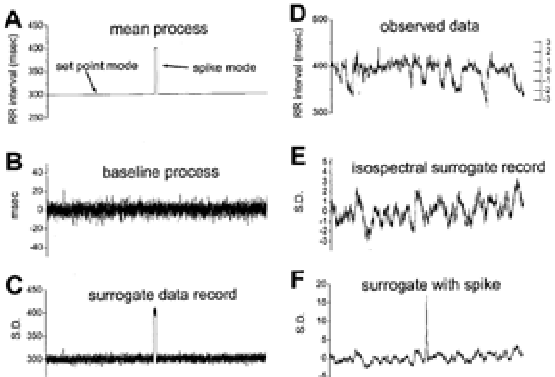
Khandoker, A., Palaniswami, M. & Begg, R. (2008) A comparative study on approximate entropy measure and poincare plot indexes of minimum foot clearance variability in the elderly during walking. *Journal of NeuroEngineering and Rehabilitation, 5, 1, 4.*

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:
$X(i)=[x(i),x(i+1),...,x(i+m-1)]$ i=1, N-m+1 (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:
$d[X(i),X(j)] = \max_{k=0,m-1} [|x(i+k)-x(j+k)|]$ (2)
3) Define for each i, for i=1, N-m+1, let
$C_r^m(i) = V^m(i)/(N-m+1)$ (3)
where $V^m(i)$ = no. of $d[X(i),X(j)] \le r$
4) Take the natural logarithm of each $C_r^m(i)$, and average it over i as defined in step 5:
$\phi^m(r) = \frac{1}{N-m+1}\sum_{i=1}^{N-m+1} \ln C_r^m(i)$ (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:
$ApEn(m,r,N) = \phi^m(r) - \phi^{m+1}(r)$ (5)

SampEn

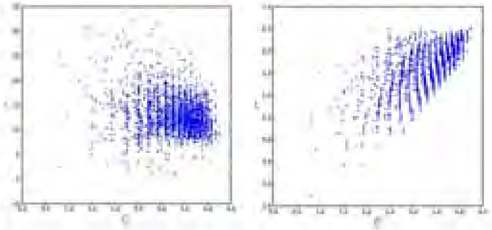
Given a signal x(n)=x(1), x(2),..., x(N), where N is the total number of data points. SampEn algorithm can be summarized as follows [7]:

1) Form m-vectors, X(1) to X(N-m+1) defined by:
$X(i)=[x(i),x(i+1),...,x(i+m-1)]$ i=1, N-m+1 (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:
$d_m[X(i),X(j)] = \max_{k=0,m-1} (|x(i+k)-x(j+k)|)$ (7)
3) Define for each i, for i=1, N-m, let
$B_i^m(r) = \frac{1}{N-m-1} \times$ no. of $d_m[X(i),X(j)] \le r, i \ne j$ (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 $d_{m+1}[X(i),X(j)] \le r, i \ne j$ (9)
5) Define $B^m(r) = \frac{1}{N-m}\sum_{i=1}^{N-m} B_i^m(r)$ (10)
$A^m(r) = \frac{1}{N-m}\sum_{i=1}^{N-m} A_i^m(r)$ (11)
6) SampEn value for a finite data length of N can be estimated:
$SampEn(m,r,N) = -\ln[A^m(r)/B^m(r)]$ (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

Advance Organizer (1/2)

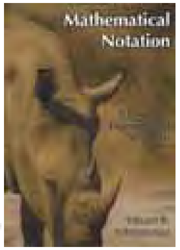
- **Abduction** = cyclical process of generating possible explanations (i.e., identification of a set of hypotheses that are able to account for the clinical case on the basis of the available data) and testing those (i.e., evaluation of each generated hypothesis on the basis of its expected consequences) 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
- **DIKW-Model** = Data-Information-Knowledge-Wisdom four 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 multi-variate
- **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

Mathematical Notation

"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
$\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$	Matrix of input samples
$\mathbf{y} = [y_1, \ldots, y_n]$	Vector of output samples
$\mathbf{Z} = [\mathbf{X}, \mathbf{y}]$	Combined input–output training data or
$\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n]$	Representation of data points in a feature space

Distribution

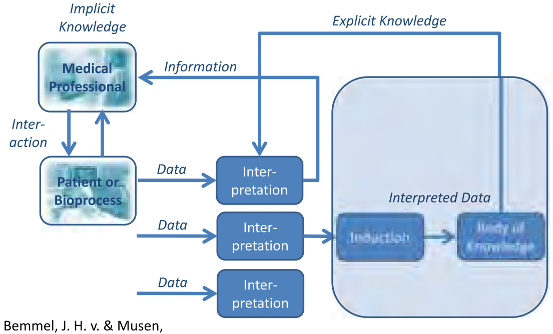
P	Probability
$F(\mathbf{x})$	Cumulative probability distribution function (cdf)
$p(\mathbf{x})$	Probability density function (pdf)
$p(\mathbf{x}, y)$	Joint probability density function
$p(\mathbf{x}; \omega)$	Probability density function, which is parameterized
$p(y\|\mathbf{x})$	Conditional density
$t(\mathbf{x})$	Target function

Glossary

- ApEn = Approximate Entropy;
- $\mathbb{C}_{data}$ = Data in computational space;
- DIK = Data-Information-Knowledge-3-Level Model;
- DIKW = Data-Information-Knowledge-Wisdom-4-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);
- $\mathbb{P}_{data}$ = Data in perceptual space;
- PDB = Protein Data Base;
- SampEn = Sample Entropy;

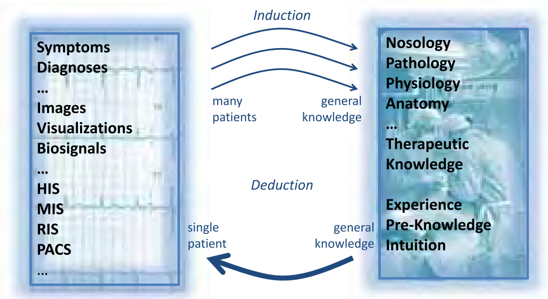
Clinical view on data – information, and knowledge

Clinical View of Data, Information, Knowledge



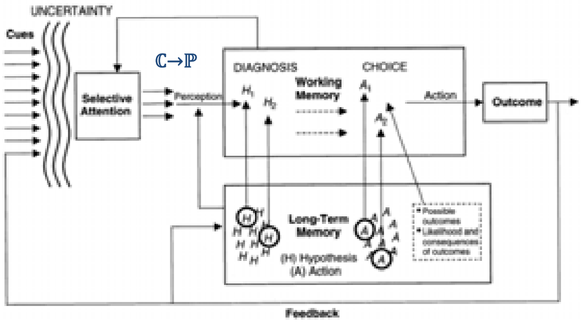
Bemmel, J. H. v. & Musen, M. A. (1997) *Handbook of Medical Informatics. Heidelberg, Springer.*

From Patient Data to Medical Knowledge



Holzinger (2007)

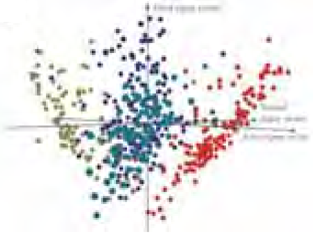
Human Information Processing Model



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

A space as a set of points

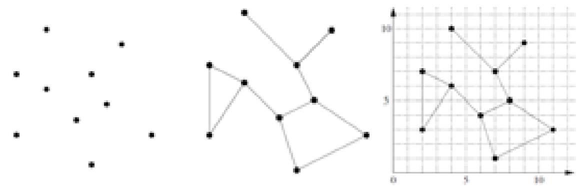
$$f : X \rightarrow \mathbb{R}$$



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.*

Point Cloud Data Sets

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

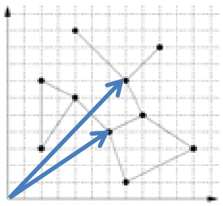


Point cloud in $\mathbb{R}^2$ topological space metric space

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

Example Metric Space

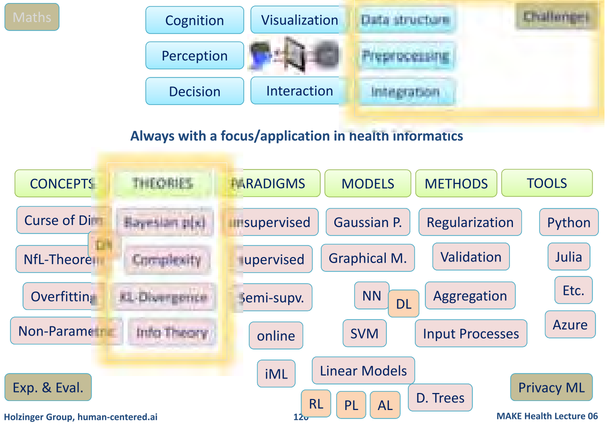
A set S with a metric function d is a metric space



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

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

ML-Jungle Top Level View and the focus of today …



Maths

Cognition Visualization Data structure Challenges

Perception Preprocessing

Decision Interaction Integration

Always with a focus/application in health informatics

CONCEPTS THEORIES PARADIGMS MODELS METHODS TOOLS

Curse of Dim Bayesian p(x) unsupervised Gaussian P. Regularization Python

NfL-Theorem DM Complexity supervised Graphical M. Validation Julia

Overfitting KL-Divergence Semi-supv. NN DL Aggregation Etc.

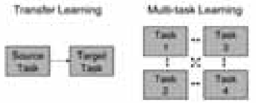
Non-Parametric Info Theory online SVM Input Processes Azure

iML Linear Models

Exp. & Eval. RL PL AL D. Trees Privacy ML

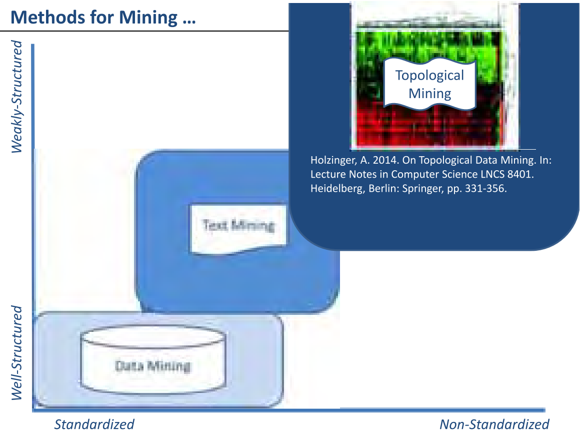
Grand challenges in Machine Learning for Health

- Big data with many training sets (this is good for ML!)
- **Small number of data sets, rare events**
- **Very-high-dimensional problems**
- **Complex data – NP-hard problems**
- **Missing, dirty, wrong, noisy, …, data**
- **GENERALISATION**
- **TRANSFER**



Torrey, L. & Shavlik, J. 2009. Transfer learning. Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques, 242-264, doi:10.4018/978-1-60566-766-9.ch011.

Methods for Mining …



Holzinger, A. 2014. On Topological Data Mining. In: Lecture Notes in Computer Science LNCS 8401. Heidelberg, Berlin: Springer, pp. 331-356.

Discrete versus continuous random variable

- X: S → ℝ (“measure” of outcome)
- Events can be defined according to X
  - E(X=a) = {s$_i$|X(s$_i$)=a}
  - E(X≥a) = {s$_i$|X(s$_i$) ≥ a}
- Consequently, probabilities can be defined on X
  - P(X=a) = P(E(X=a))
  - P(a≥X) = P(E(a≥X))
- **partitioning the sample space**

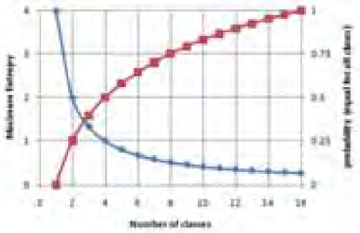
εντροπια

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.*

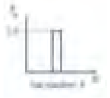
A measure for uncertainty (2/3)

$$\log_2 \frac{1}{p} = -\log_2 p \qquad H = -\sum_{i=1}^{N} p_i log_2(p_i)$$



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

Entropy H as a measure for uncertainty (3/3)



$$H_B = -\sum_{k=1} p_k \log_2 p_k = -1 * \log_2(1) = 0$$

$$H_B = -\sum_{k=1}^{B} \frac{1}{B} \log_2 \frac{1}{B} = \log_2(B)$$

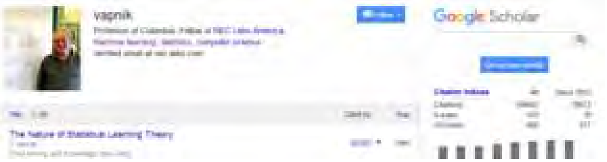
$$H = H_{max} = \log_2 N$$

Background on Information Theory

- Developed by Claude Shannon in the 1940s
- Maximizing the amount of information that can be transmitted over an imperfect communication channel
- Data compression (entropy)
- Transmission rate (channel capacity)

Claude E. Shannon: A Mathematical Theory of Communication, Bell System Technical Journal, Vol. 27, pp. 379–423, 623–656, 1948

Vapnik – Chervonenkis



vapnik

The Nature of Statistical Learning Theory

- The VC dimension is a measure of the capacity of a space of functions that can be learned by a statistical classification algorithm. It is defined as the cardinality of the largest set of points that the algorithm can shatter. It is a core concept in Vapnik–Chervonenkis theory

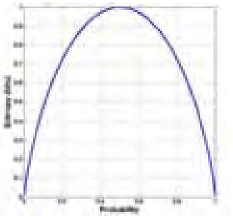
Vapnik, V. N. & Chervonenkis, A. Y. 1971. On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities. Theory of Probability & Its Applications, 16, (2), 264-280, doi:10.1137/1116025.

Entropy H as a measure for uncertainty (1/3)

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

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

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



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.*

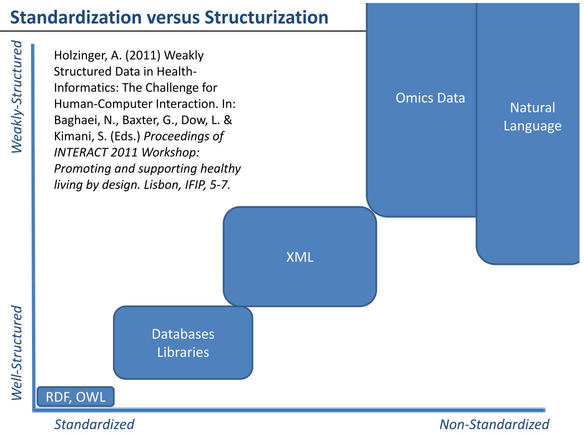
Entropic methods – what for?

- 1) Set of noisy, complex 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.

Standardization versus 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.*



Weakly-Structured

Well-Structured

Standardized

Non-Standardized

Omics Data

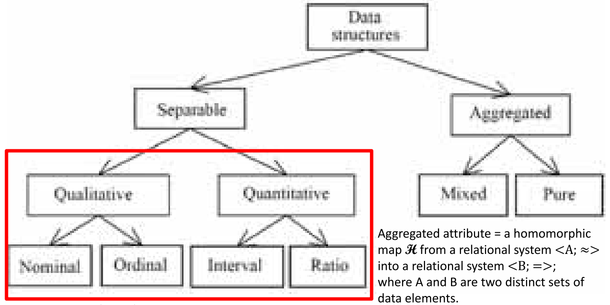
Natural Language

XML

Databases Libraries

RDF, OWL

Example: Data structures - Classification



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

Aggregated attribute = a homomorphic map 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.

Categorization of Data (Classic “scales”)

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

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