Andreas Holzinger

VO 709.049 Medical Informatics

21.10.2015 11:15-12:45

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

a.holzinger@tugraz.at

Tutor: markus.plass@student.tugraz.at

Web: 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

- Computational space (high-dimensional)
- Data structures
- DIK-Model
- DIKW-Model
- Dimensionality of data
- Information complexity
- Information entropy
- Perceptual space (low-dimensional)
- Standardization versus Structurization

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

Advance Organizer (1/2)

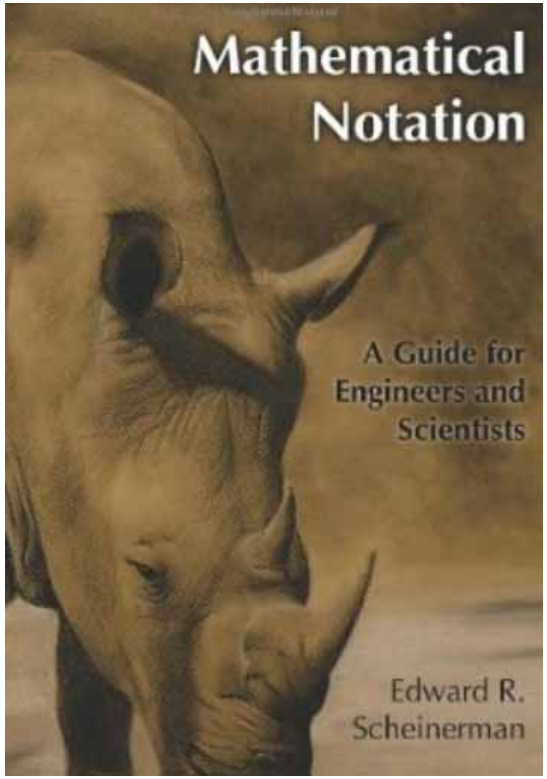
- **Abduction** = cyclical process of generating possible explanations (i.e., identification of a set of hypotheses that are able to account for the clinical case on the basis of the available data) and testing those (i.e., evaluation of each generated hypothesis on the basis of its expected consequences)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

Common Mathematical Notations with LaTeX commands

"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;

Key Problems

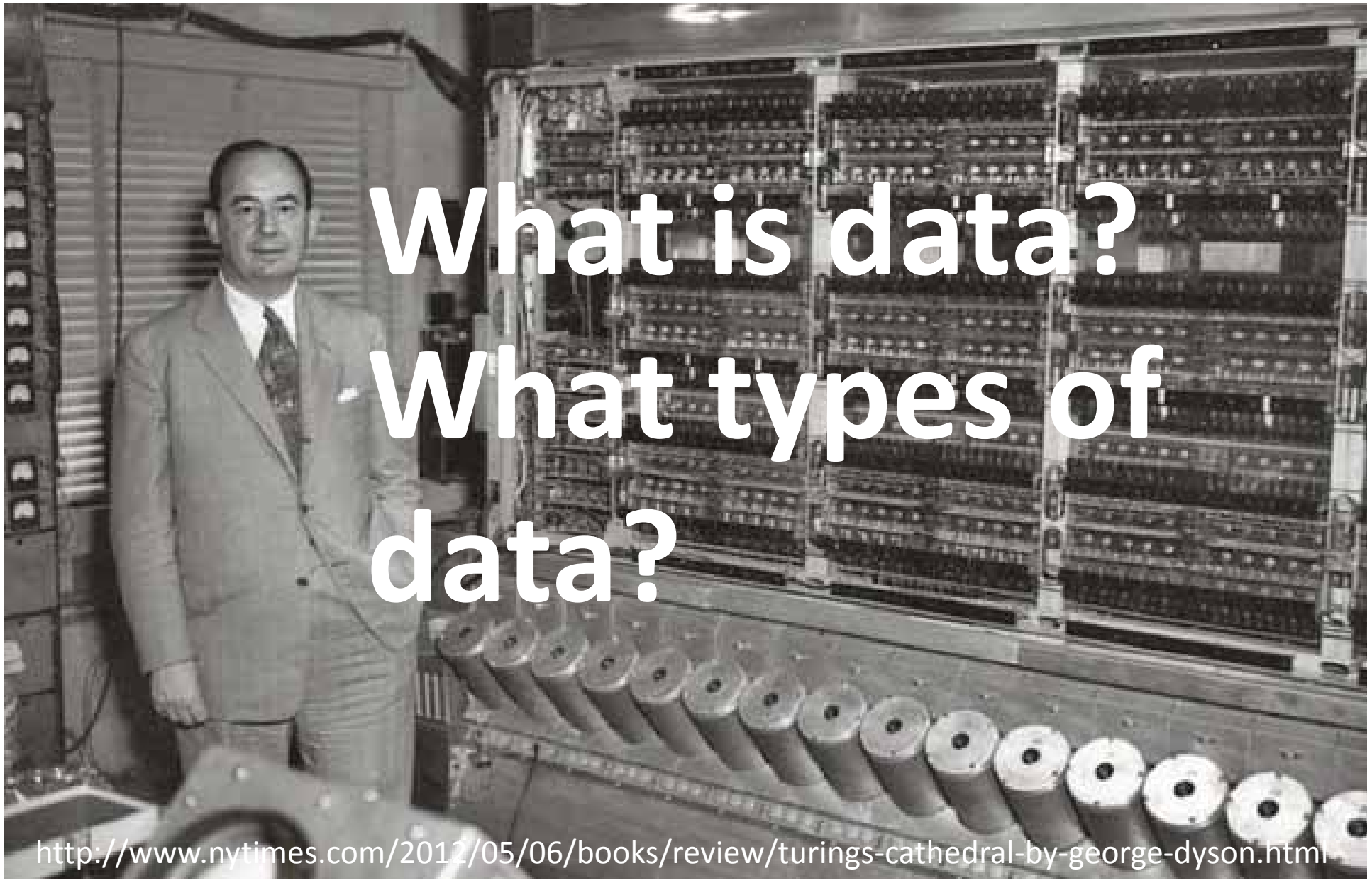
- Heterogeneous, distributed, inconsistent data sources (need for **data integration** & fusion) [1]
- **Complex data** (high-dimensionality – challenge of dimensionality reduction and visualization) [2]
- Noisy, uncertain, missing, dirty, and imprecise, imbalanced data (challenge of **pre-processing**)
- 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.

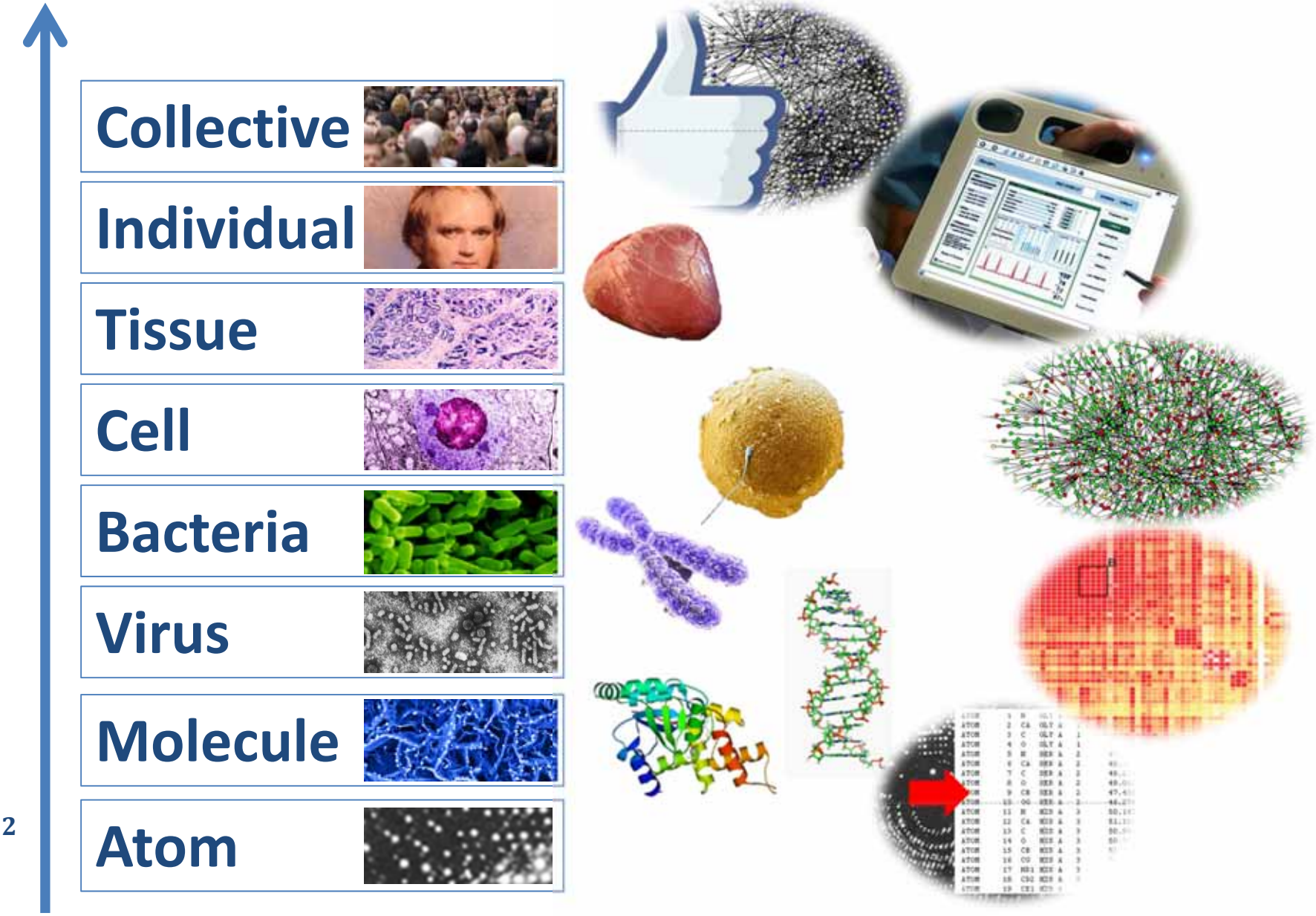
- Data in traditional Statistics
- Low-dimensional data ($< \mathbb{R}^{100}$)
- Problem: Much noise in the data
- Not much structure in the data but it can be represented by a simple model

- 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.


What is data?
What types of data?
http://www.nytimes.com/2012/05/06/books/review/turings-cathedral-by-george-dyson.html

Slide 2-1: Biomedical Data Sources

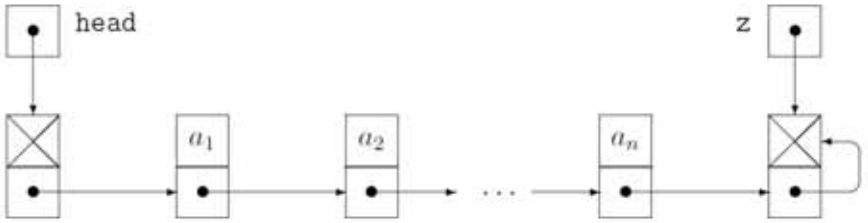


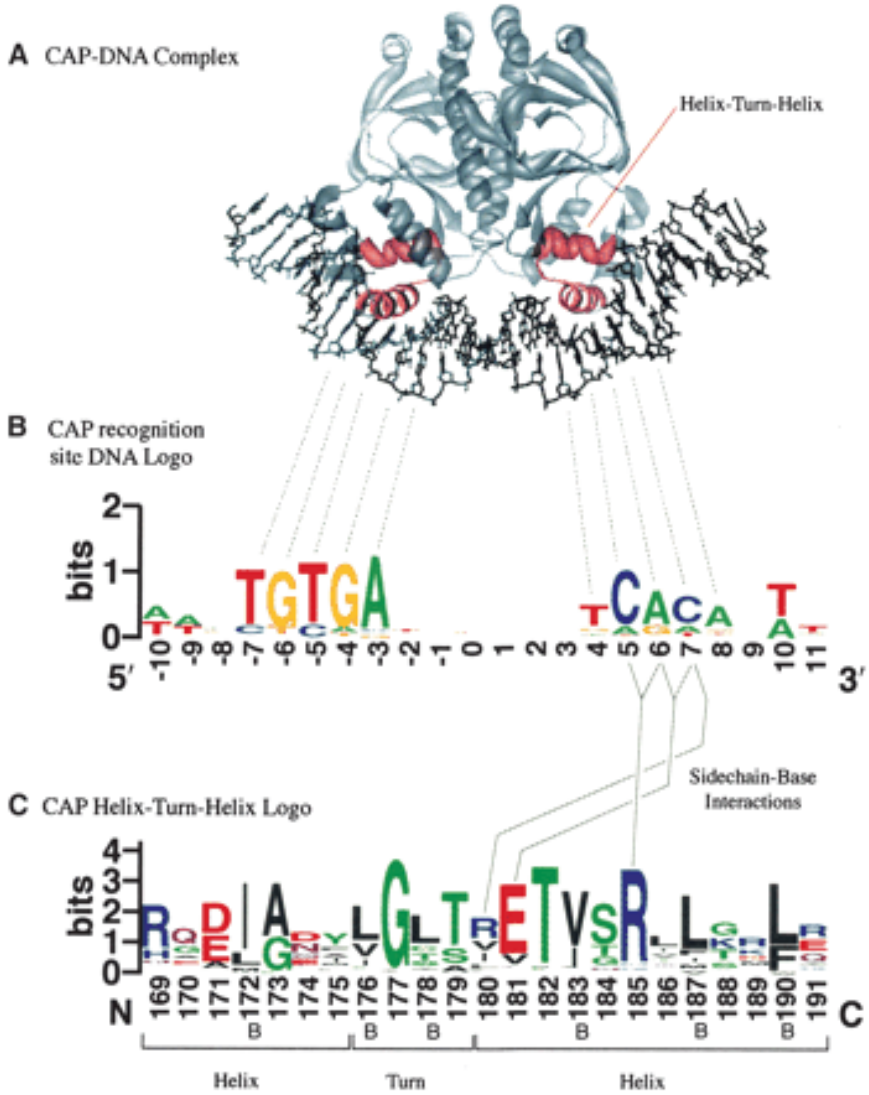
- **Physical level** -> bit = binary digit = **b**asic **i**ndissoluble uni**t** (= Shannon, Sh), ≠ Bit (!)
  in Quantum Systems -> qubit
- **Logical Level** -> integers, booleans, characters, floating-point numbers, alphanumeric strings, …
- **Conceptual (Abstract) Level** -> data-structures, e.g. lists,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

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

`TYPE link = REF node ;`
`node = RECORD`
`key : ItemType;`
`next : link;`
`END;`
	key
	next
		`class link {`
		`ItemType key;`
		`link next;`
		`}`

`VAR p, q : link ;`	p q	`link p,q;`
`p := NEW(link);`	p q	`p=new link();`
`p^.key:=x;`	p q x	`p.key=x;`
`q := NEW(link) ;`	p q x	`q=new link();`

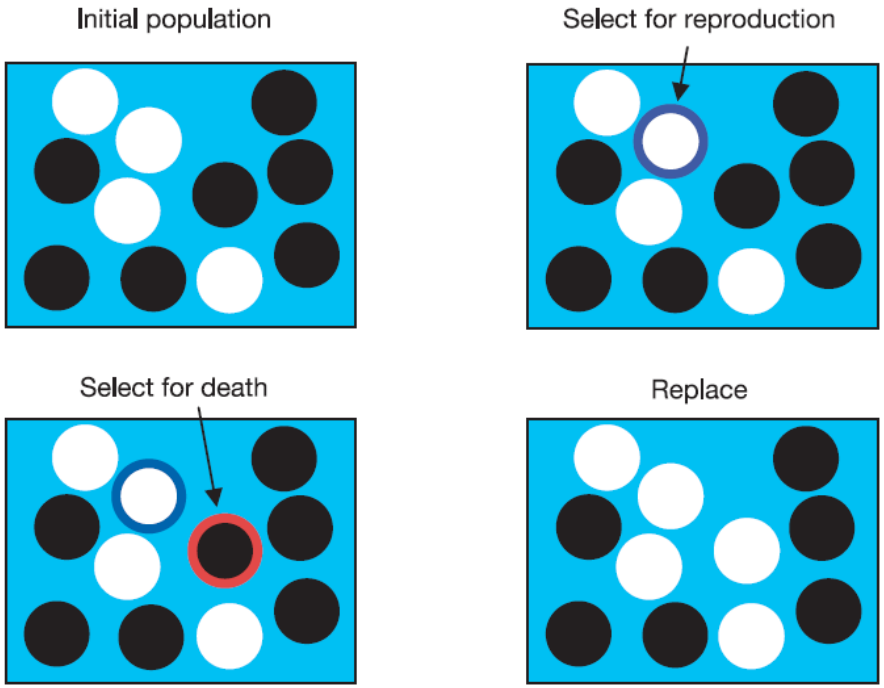


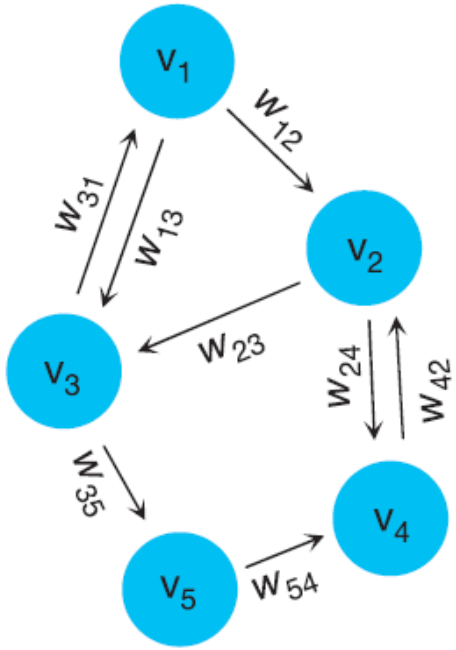


Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. (2004) WebLogo: A sequence logo generator. Genome Research, 14, 6, 1188-1190.

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

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



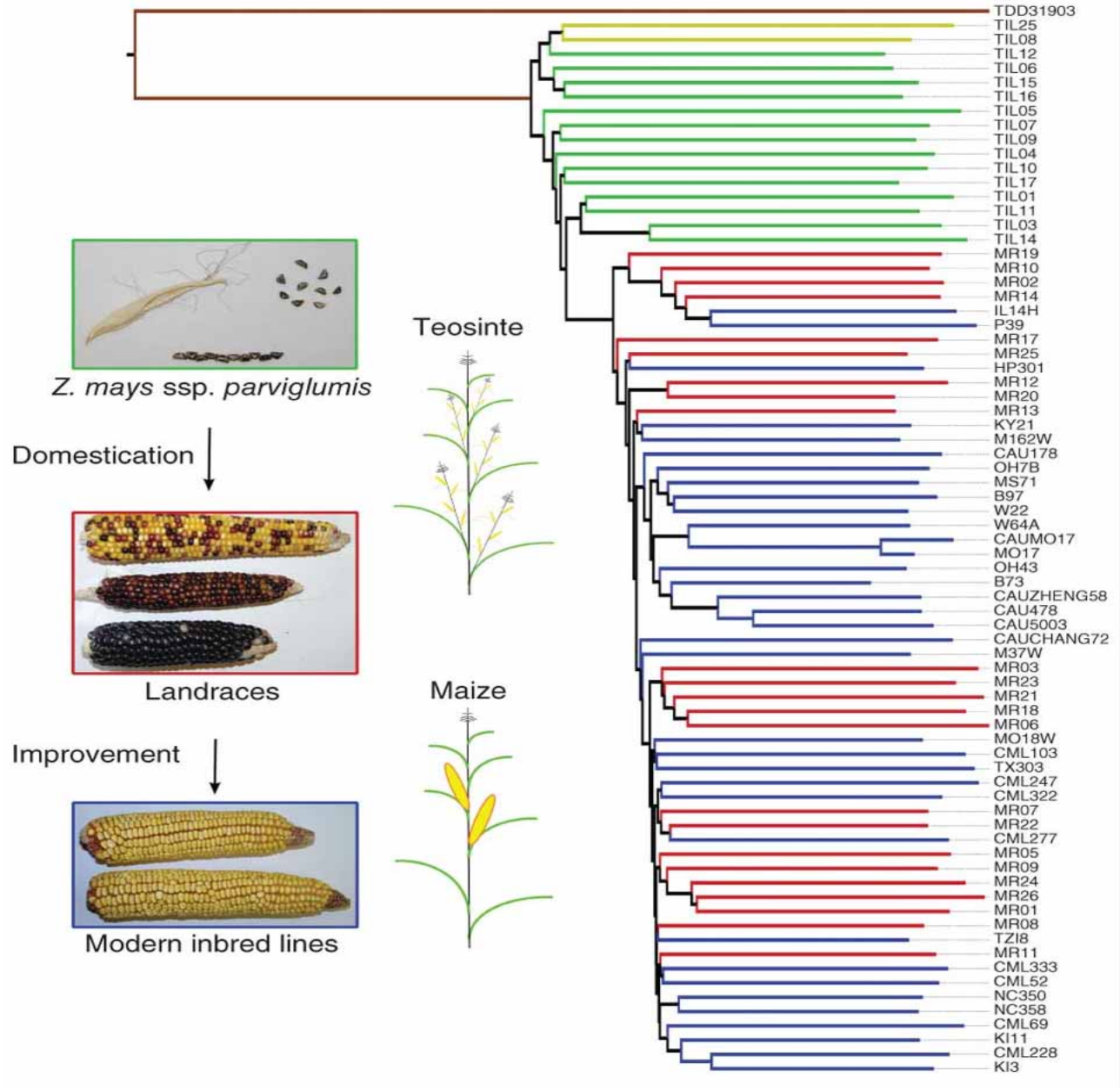


$$W = \begin{bmatrix} 0 & w_{12} & w_{13} & 0 & 0 \\ 0 & 0 & w_{23} & w_{24} & 0 \\ w_{31} & 0 & 0 & 0 & w_{35} \\ 0 & w_{42} & 0 & 0 & 0 \\ 0 & 0 & 0 & w_{54} & 0 \end{bmatrix}$$

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

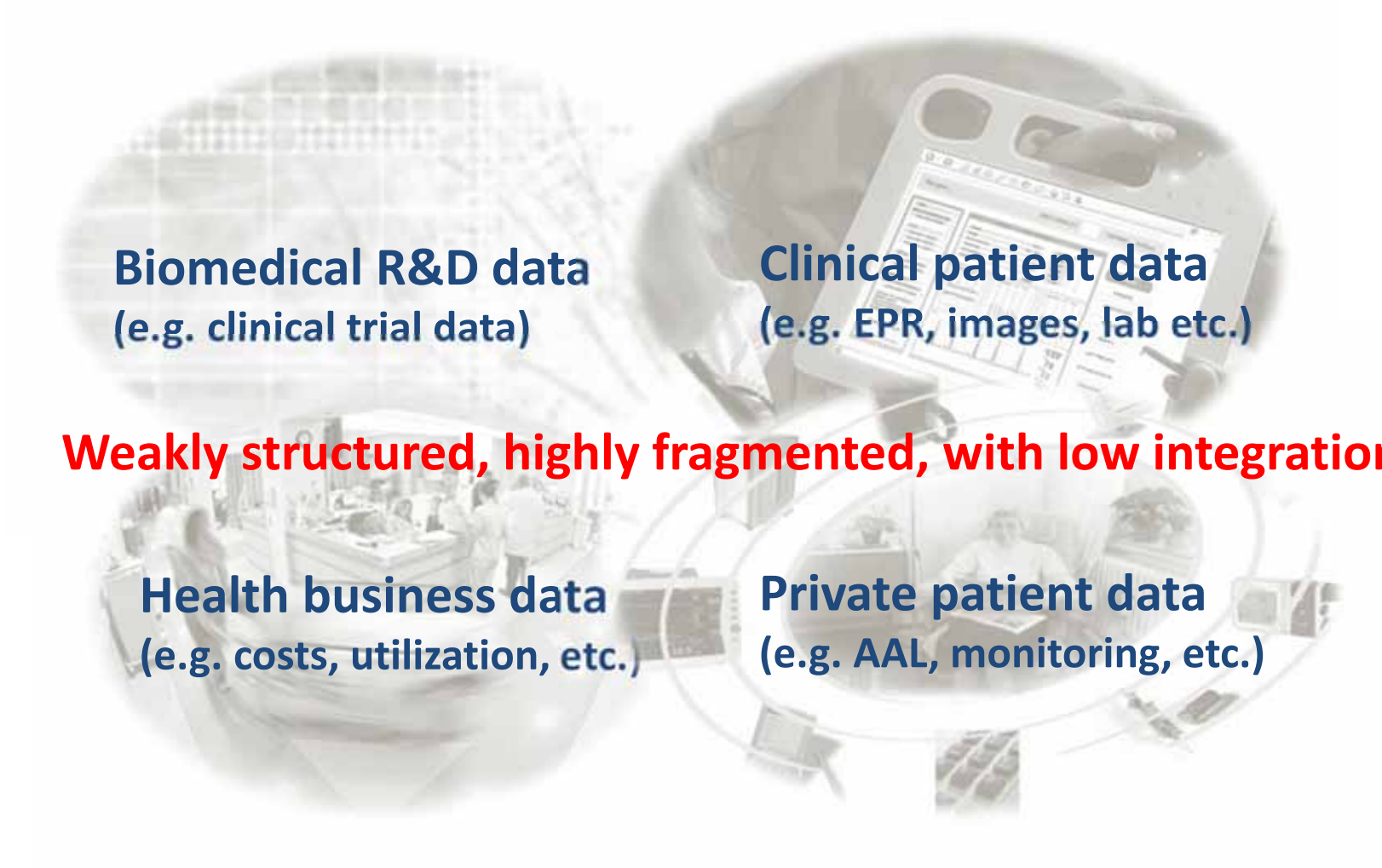
Slide 2-5: Example Data Structures (3/3) Tree

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



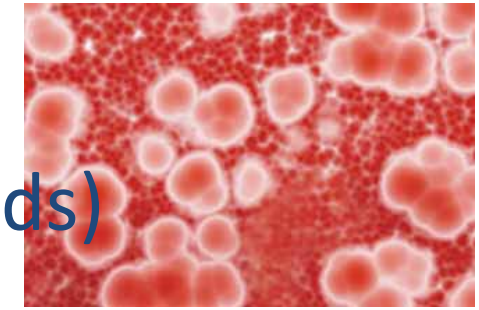
Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. & Byers, A. H. (2011) *Big data: The next frontier for innovation, competition, and productivity. Washington (DC), McKinsey Global Institute.*

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[74]	• SNP effect on protein activity or abundance	• Enzyme annotation	• Binding-site identification[75]	• Functional annotation[79]	• Functional annotation	• Functional annotation[71,103] • Biomarkers[125]
	Transcriptomics (microarray, SAGE)	• Protein: transcript correlation[20]	• Enzyme annotation[109]	• Gene-regulatory networks[76]	• Functional annotation[89] • Protein complex identification[82]		• Functional annotation[102]
		Proteomics (abundance, post-translational modification)	• Enzyme annotation[99]	• Regulatory complex identification	• Differential complex formation	• Enzyme capacity	• Functional annotation
			Metabolomics (metabolite abundance)	• Metabolic-transcriptional response		• Metabolic pathway bottlenecks	• Metabolic flexibility • Metabolic engineering[109]
				Protein–DNA interactions (ChIP–chip)	• Signalling cascades[89,102]		• Dynamic network responses[84]
					Protein–protein interactions (yeast 2H, coAP–MS)		• Pathway identification activity[89]
						Fluxomics (isotopic tracing)	• Metabolic engineering
							Phenomics (phenotype arrays, RNAi screens, synthetic lethals)

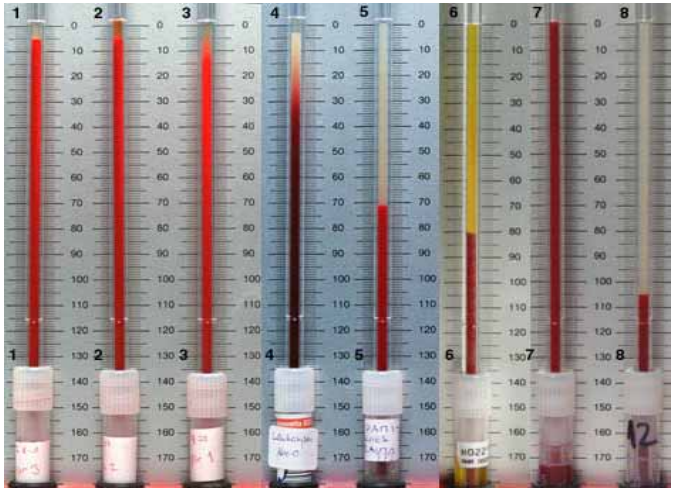
Joyce, A. R. & Palsson, B. Ø. 2006. The model organism as a system: integrating'omics' data sets. *Nature Reviews Molecular Cell Biology, 7,* ***198-210.***

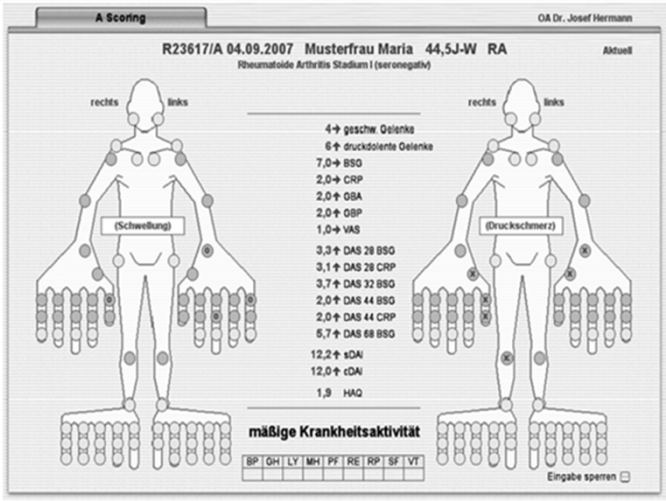
- **Gen**omics (sequence annotation)
- **Transcript**omics (microarray)
- **Prote**omics (Proteome Databases)
- **Metabol**omics (enzyme annotation)
- **Flux**omics (isotopic tracing, metabolic pathways)
- **Phen**omics (biomarkers)
- **Epigen**omics (epigenetic modifications)
- **Microbi**omics (microorganisms)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



Slide 2-8: Example of typical clinical data sets

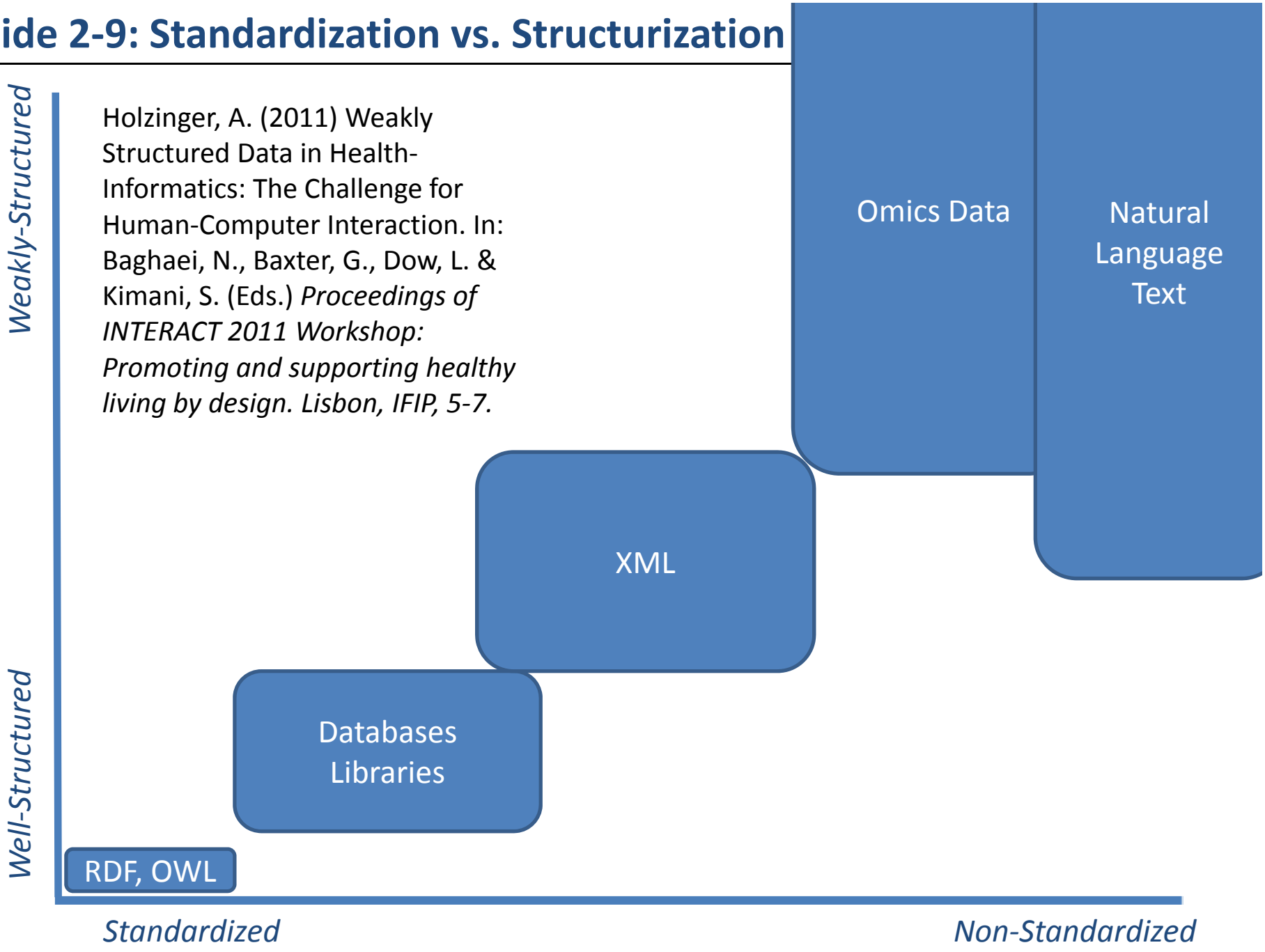
- 50+ Patients per day ~ 5000 data points per day …
- Aggregated with specific scores (Disease Activity Score, DAS)
- Current patient status is related to previous data
- = convolution over time
- ⇒ **time-series data**





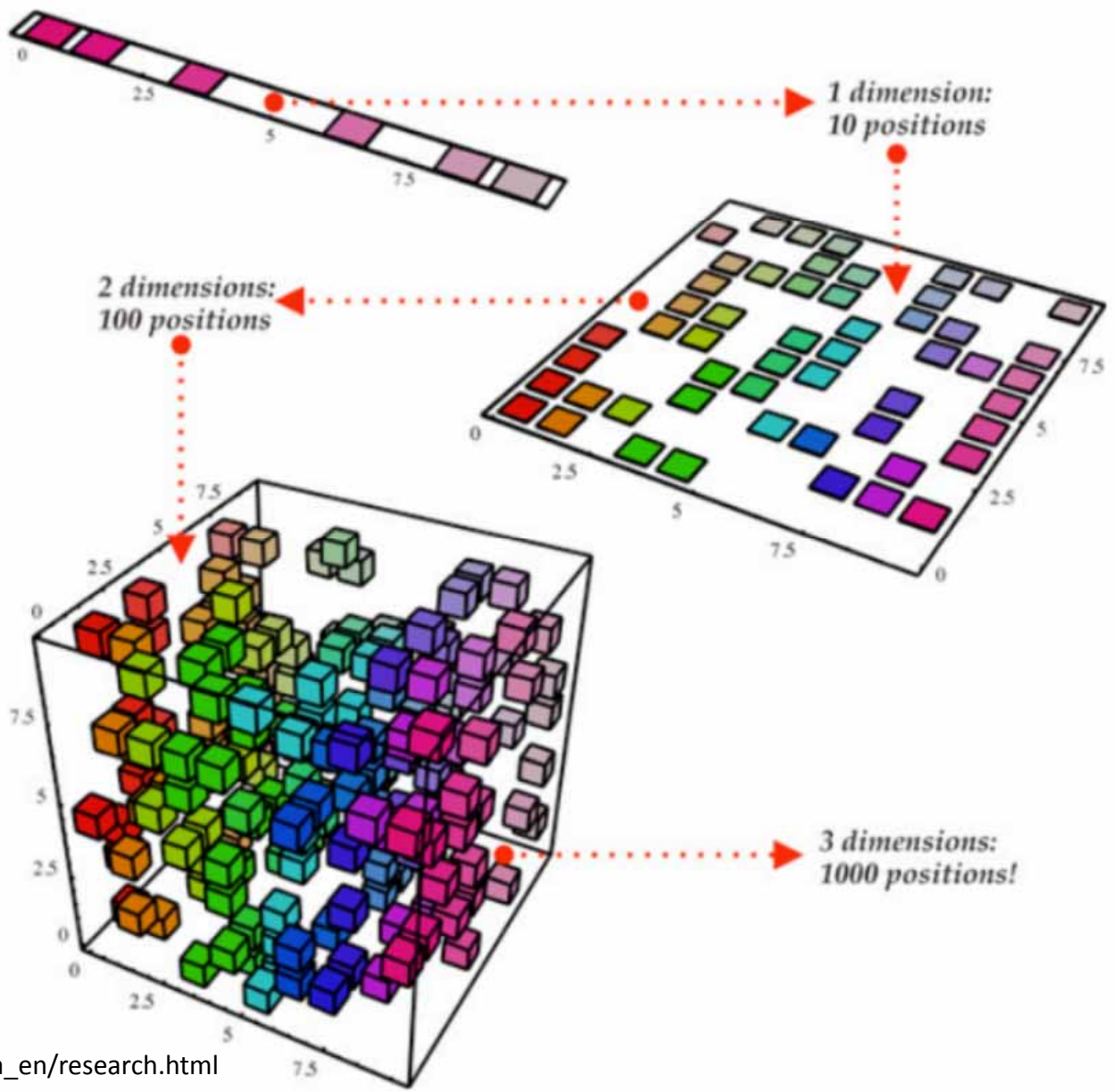
Simonic, K. M., Holzinger, A., Bloice, M. & Hermann, J. (2011). *Optimizing Long-Term Treatment of Rheumatoid Arthritis with Systematic Documentation. Pervasive Health - 5th International Conference on Pervasive Computing Technologies for Healthcare, Dublin, IEEE, 550-554.*

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



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

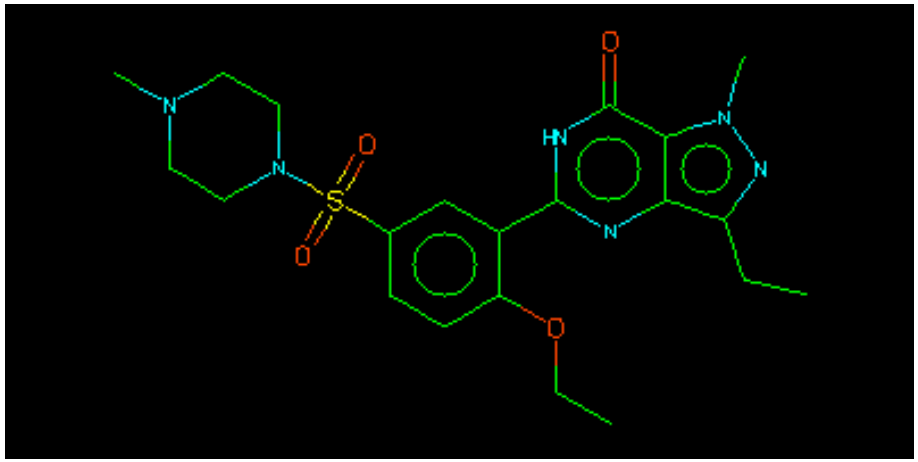
- 0-D data = a data point existing isolated from other data, e.g. integers, letters, Booleans, etc.
- 1-D data = consist of a string of 0-D data, e.g. Sequences representing nucleotide bases and amino acids, SMILES etc.
- 2-D data = having spatial component, such as images, NMR-spectra etc.
- 2.5-D data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

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



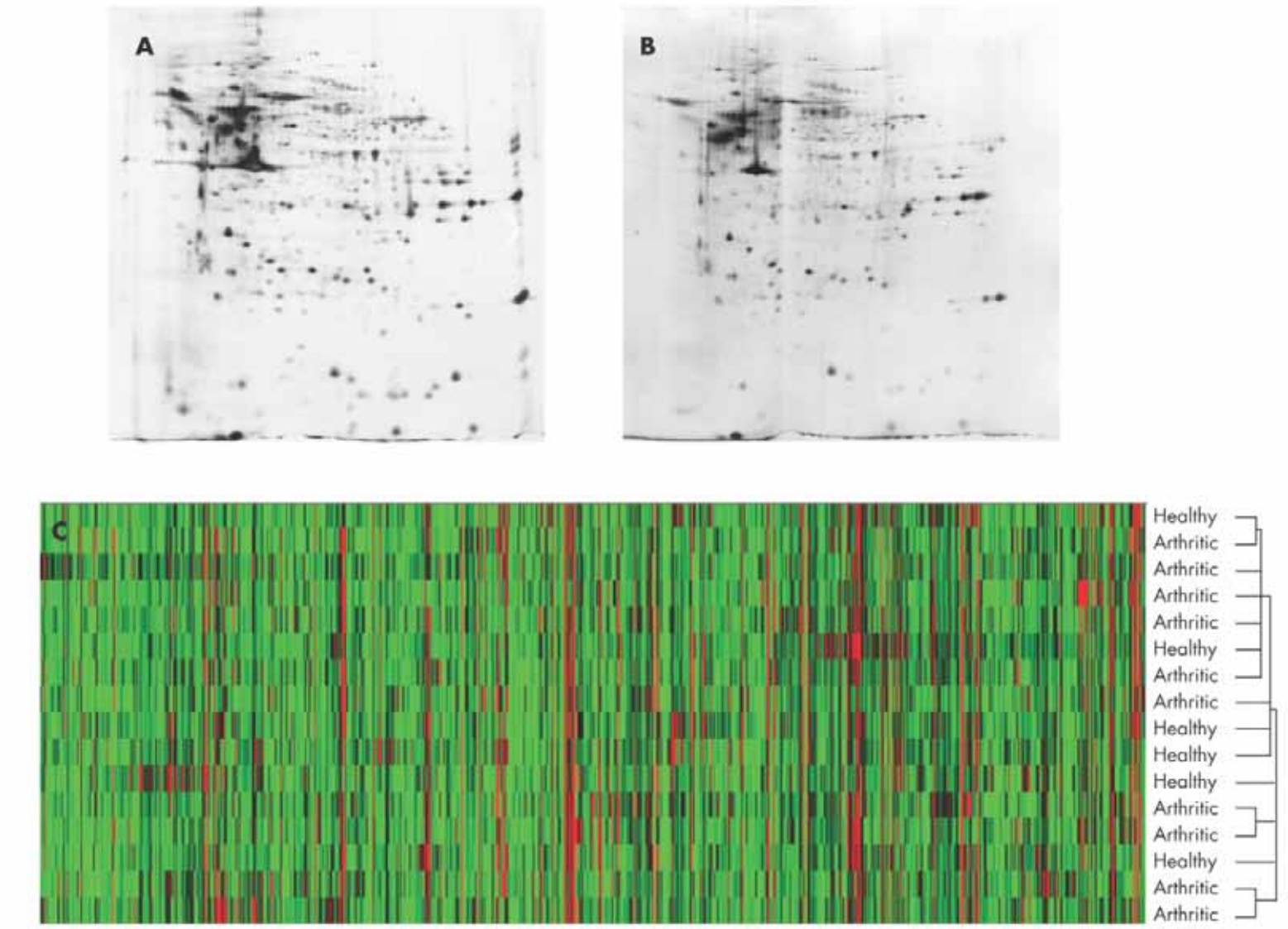
...is Canonicalizable

...is Comprehensive

...is Well Documented

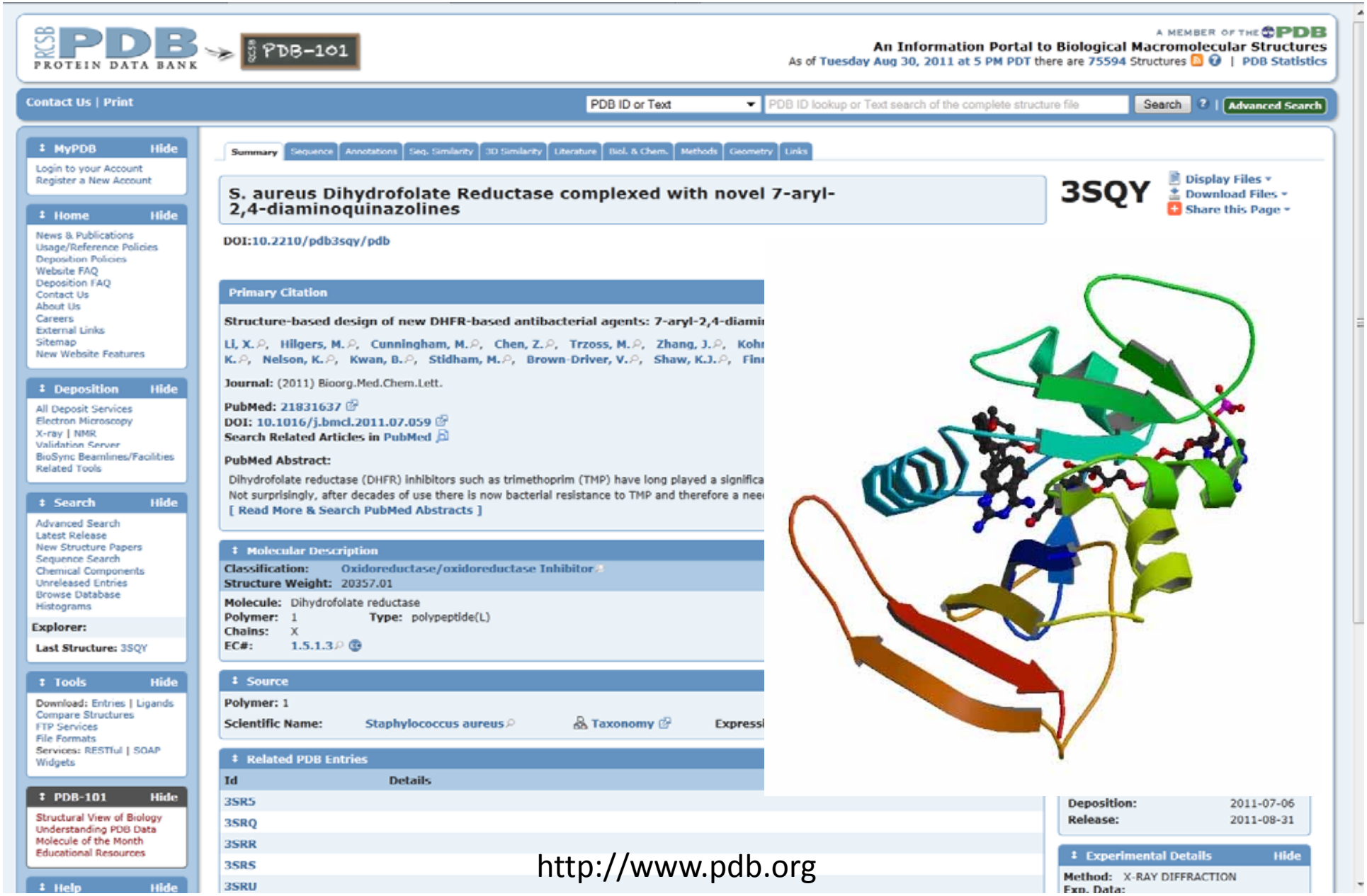
http://www.daylight.com/dayhtml_tutorials/languages/smiles/index.html

Example: 2-D data (bivariate data)



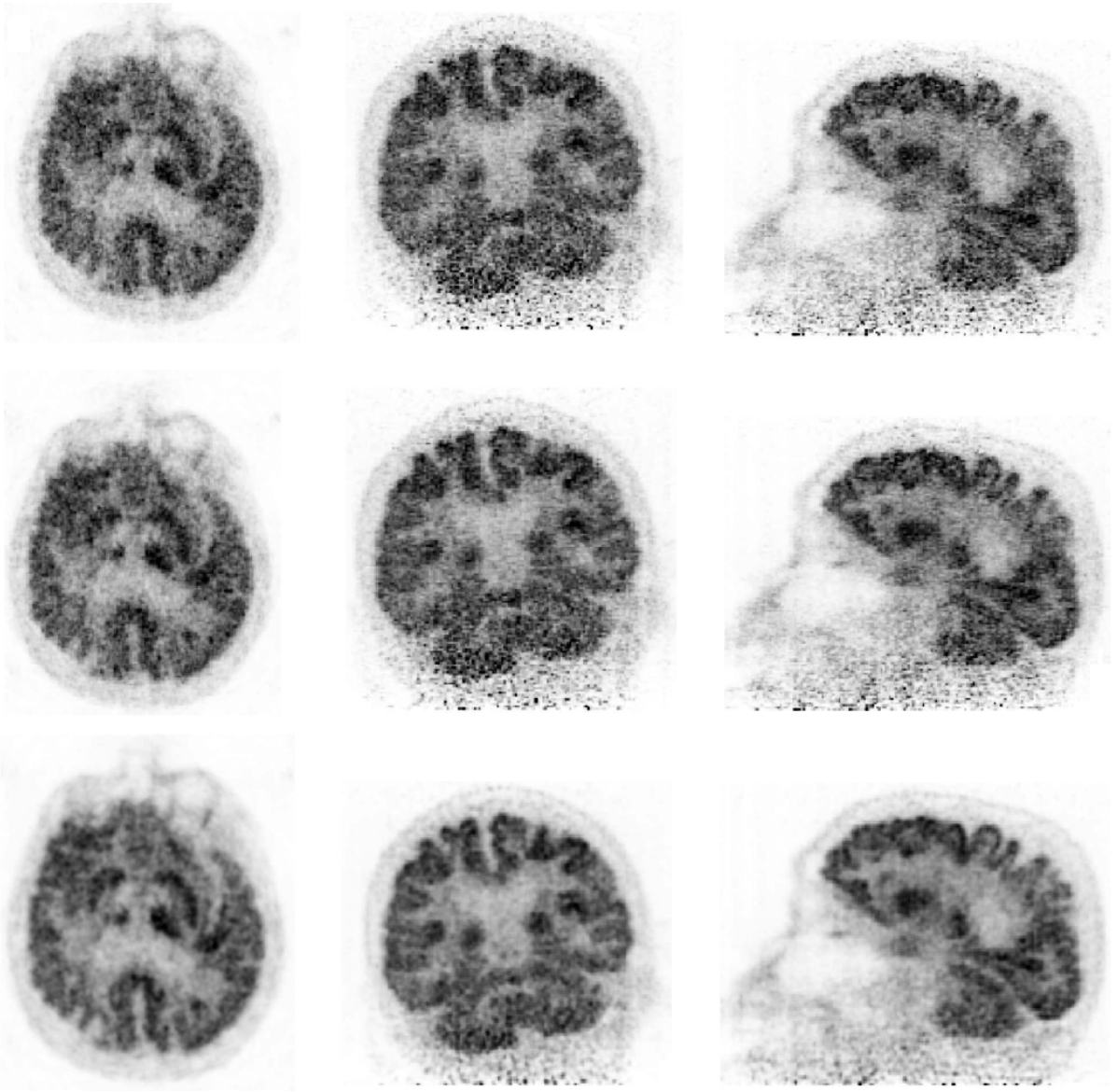
Kastrinaki et al. (2008) Functional, molecular & proteomic characterisation of bone marrow mesenchymal stem cells in rheumatoid arthritis. *Annals of Rheumatic Diseases, 67, 6, 741-749.*

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



http://www.pdb.org

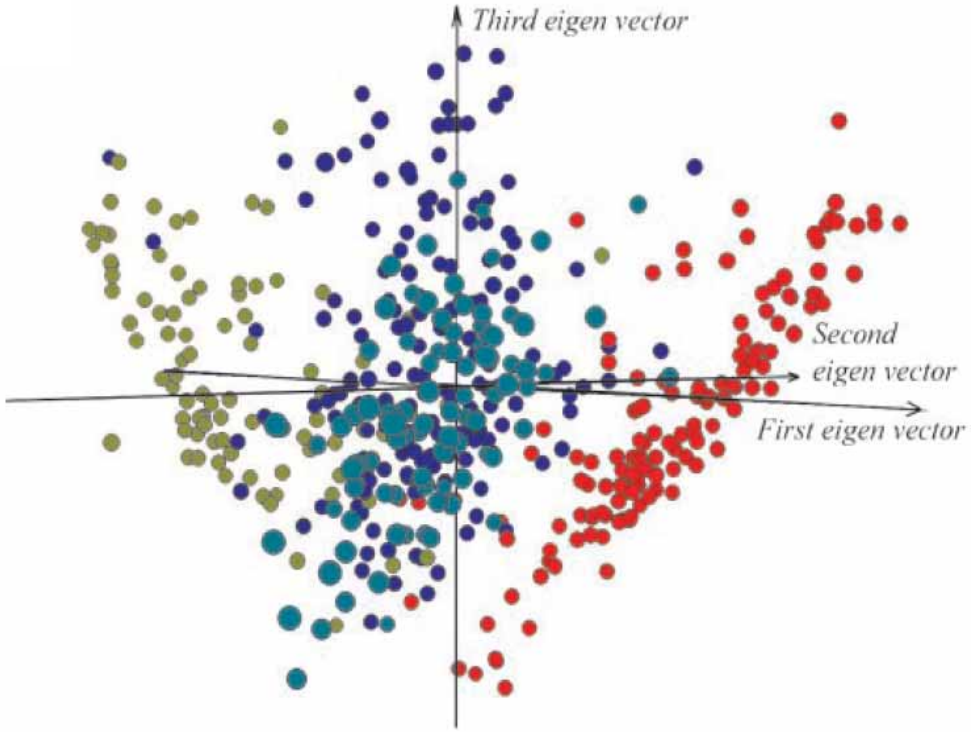
Example: 3-D Voxel data (volumetric picture elements)



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.*

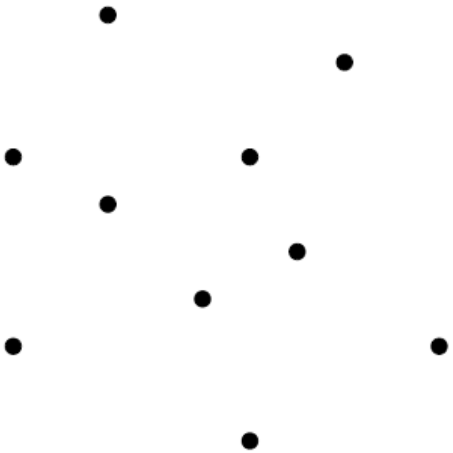
Slide 2-11 A space is a set of points

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

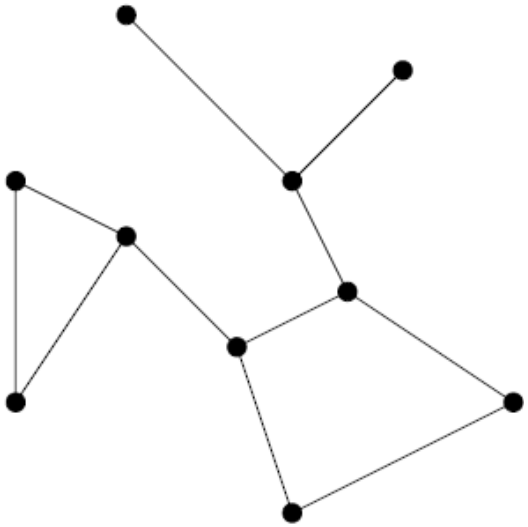


Hou, J., Sims, G. E., Zhang, C. & Kim, S.-H. 2003. A global representation of the protein fold space. *Proceedings of the National Academy of Sciences, 100, (5), 2386-2390.*

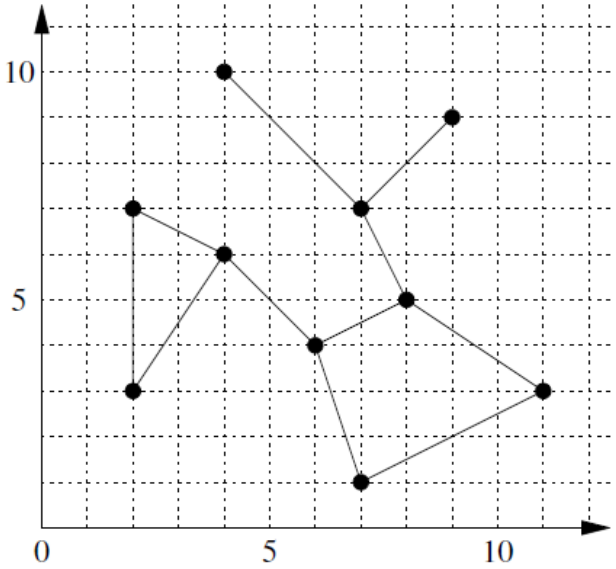
Let us collect n-dimensional i observations: $\boldsymbol{x}_i = [x_{i1}, \ldots, x_{in}]$



Point cloud in $\mathbb{R}^2$



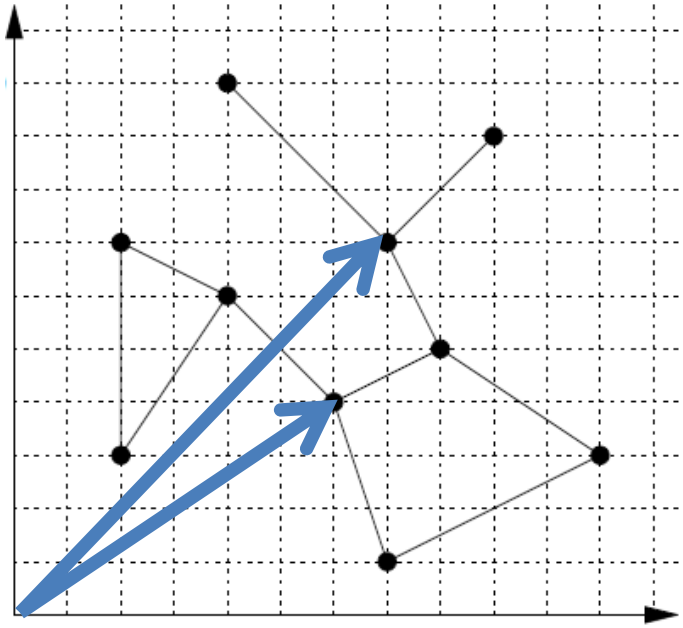
topological space



metric space

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

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.*

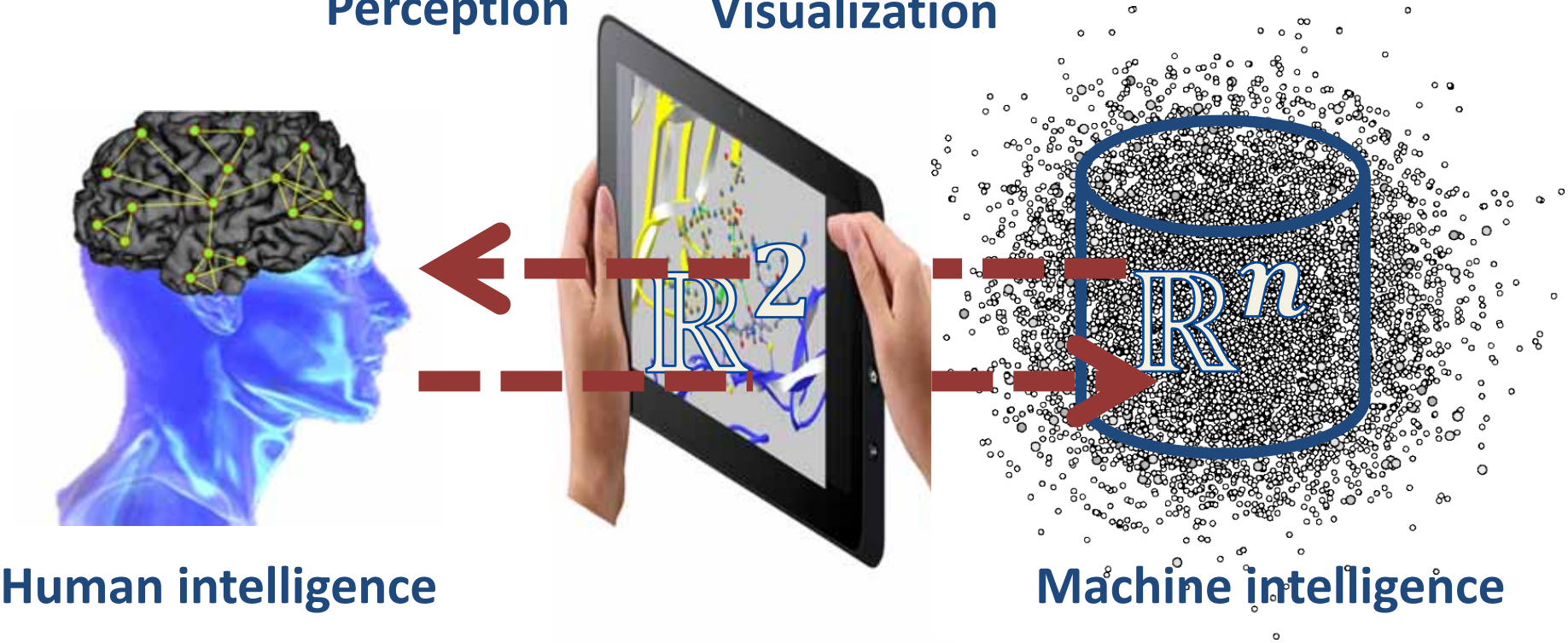
Slide 2-14 Mapping Data $\mathbb{R}^n + t \rightarrow \mathbb{R}^2 + t \; or \; \mathbb{R}^3 + t$

Cognitive Space

Computational space

Perception

Visualization



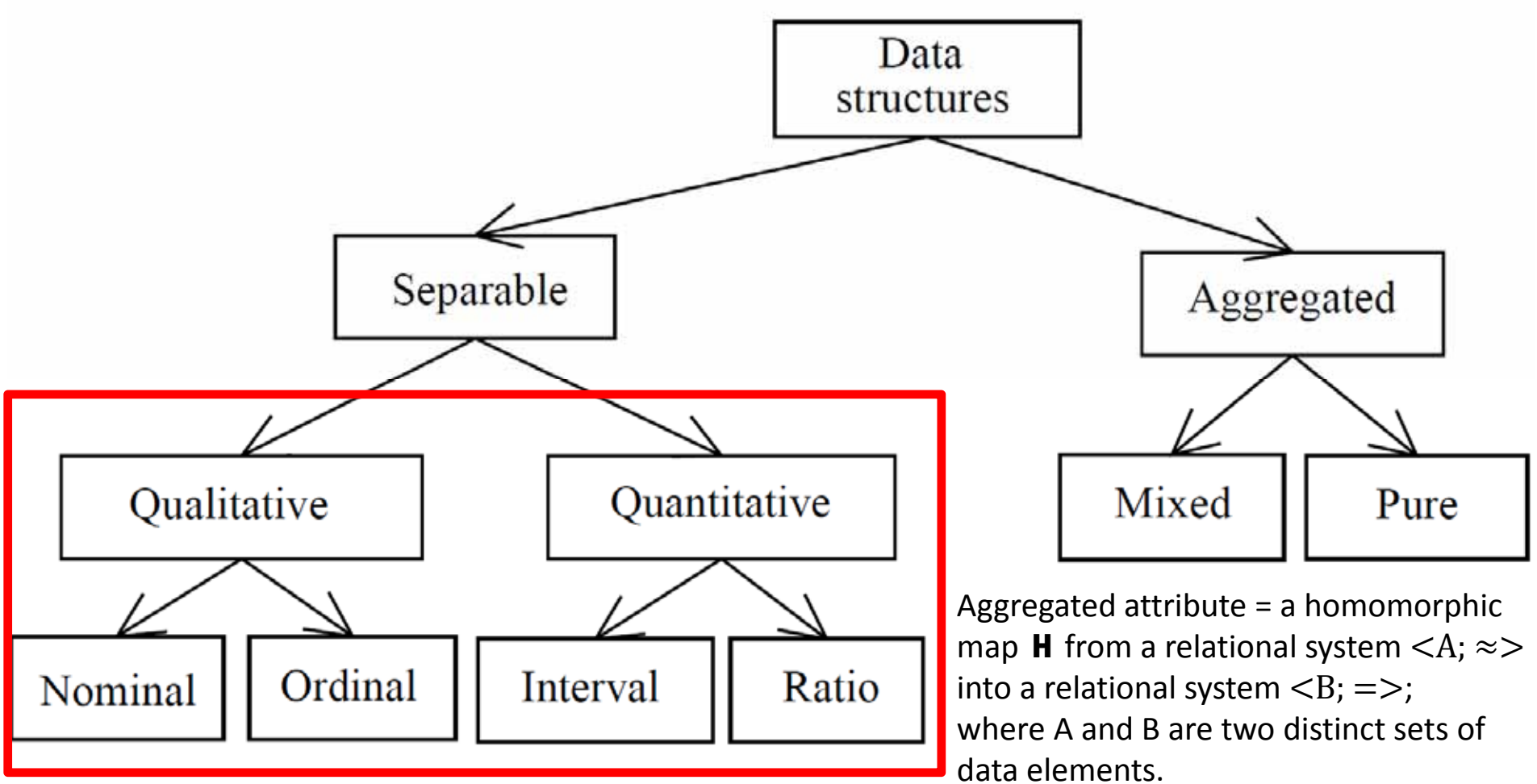
Human intelligence

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



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.

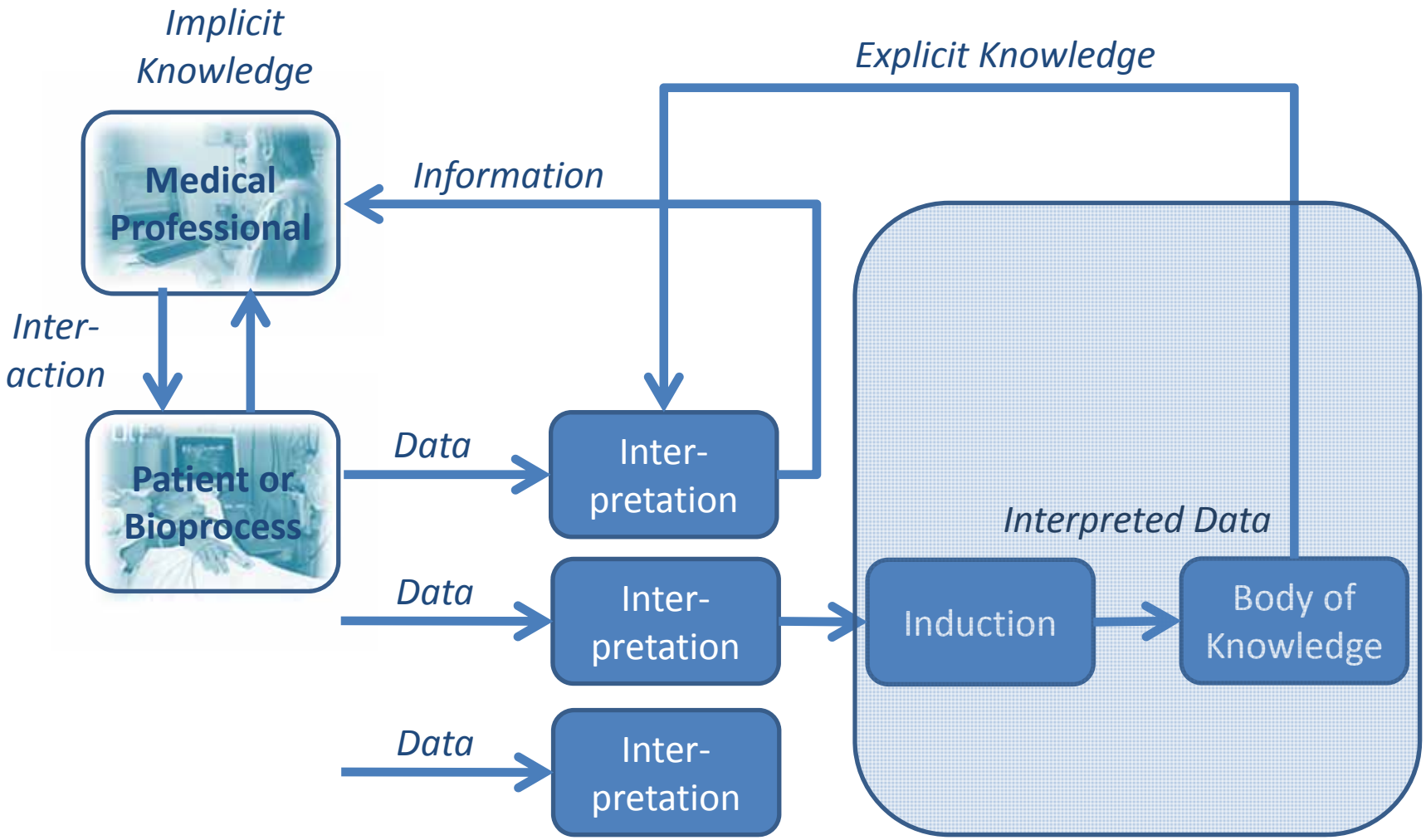
Slide 2-15: Categorization of Data (Classic "scales")

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.*

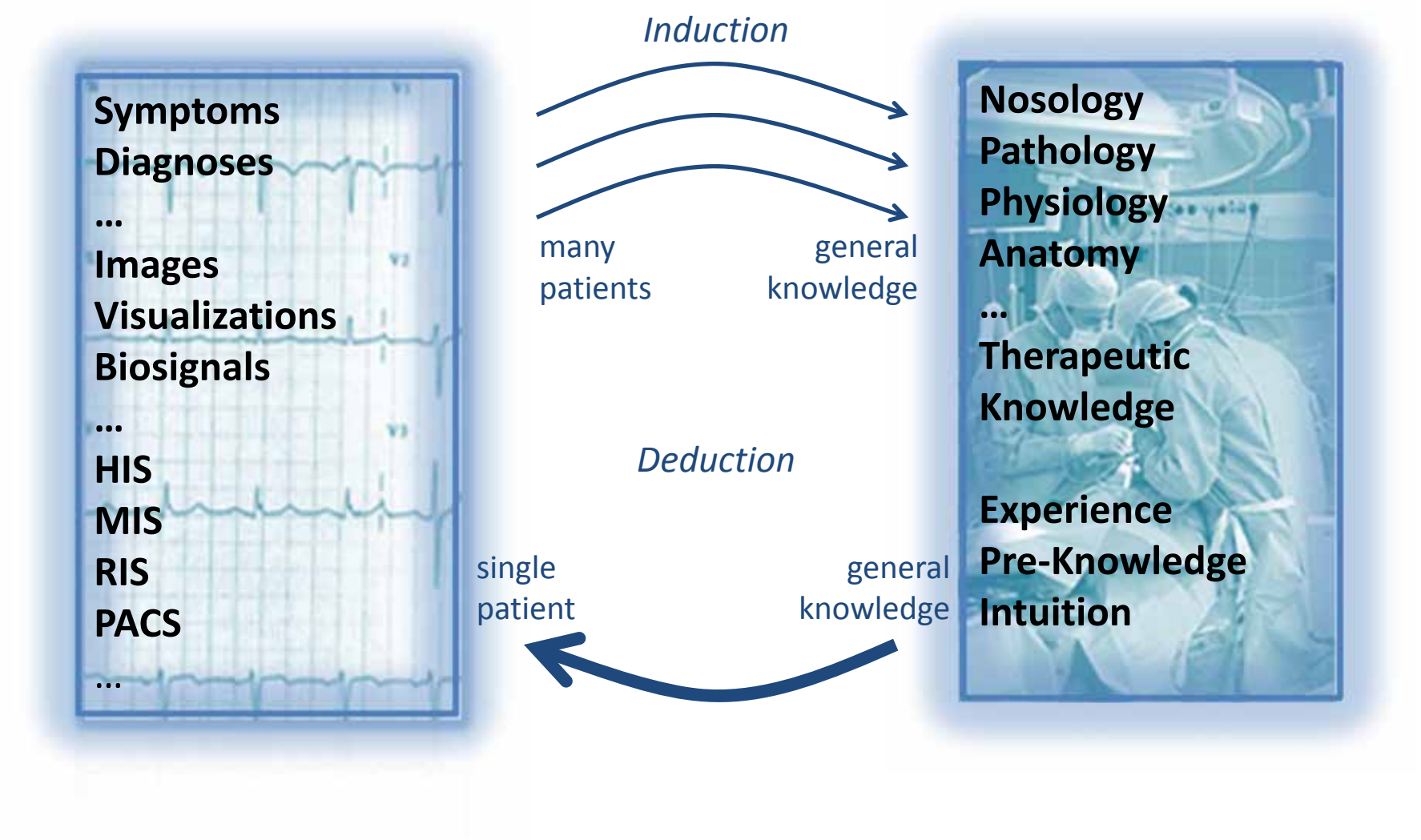

A clinical view on
data – information -
knowledge

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

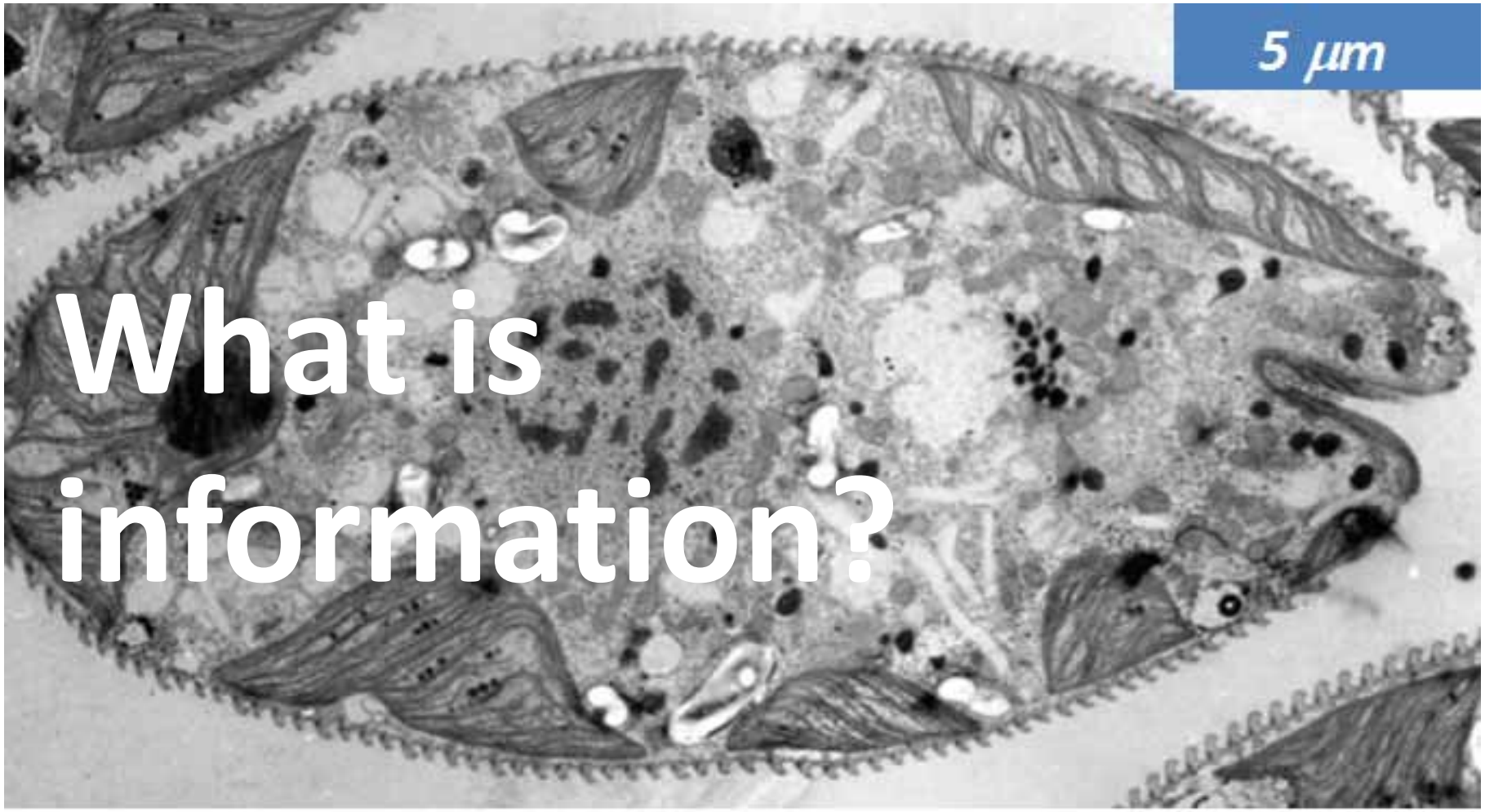


Bemmel, J. H. v. & Musen, M. A. (1997) *Handbook of Medical Informatics. Heidelberg, Springer.*

Slide 2-17: From Patient Data to Medical Knowledge

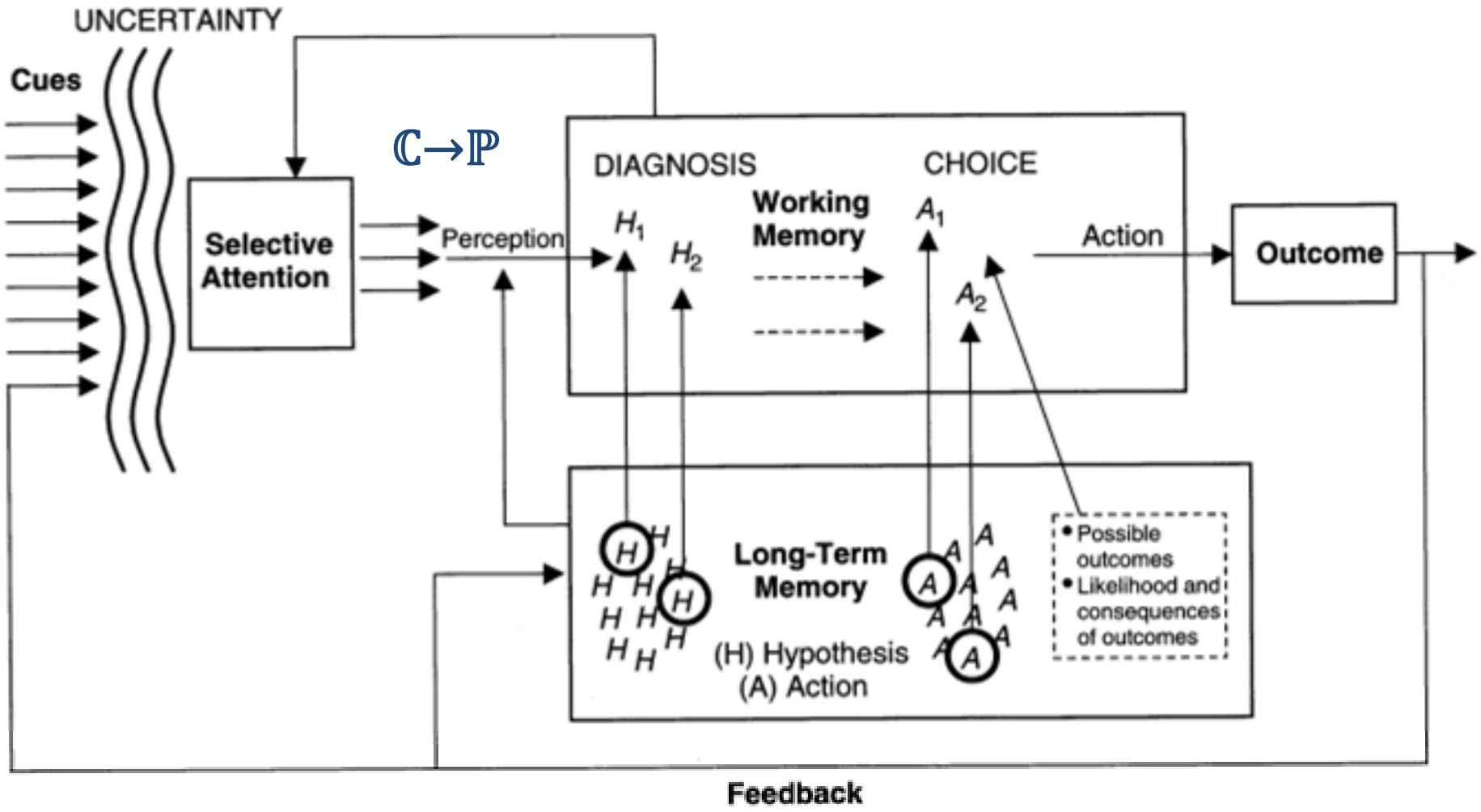


Holzinger (2007)



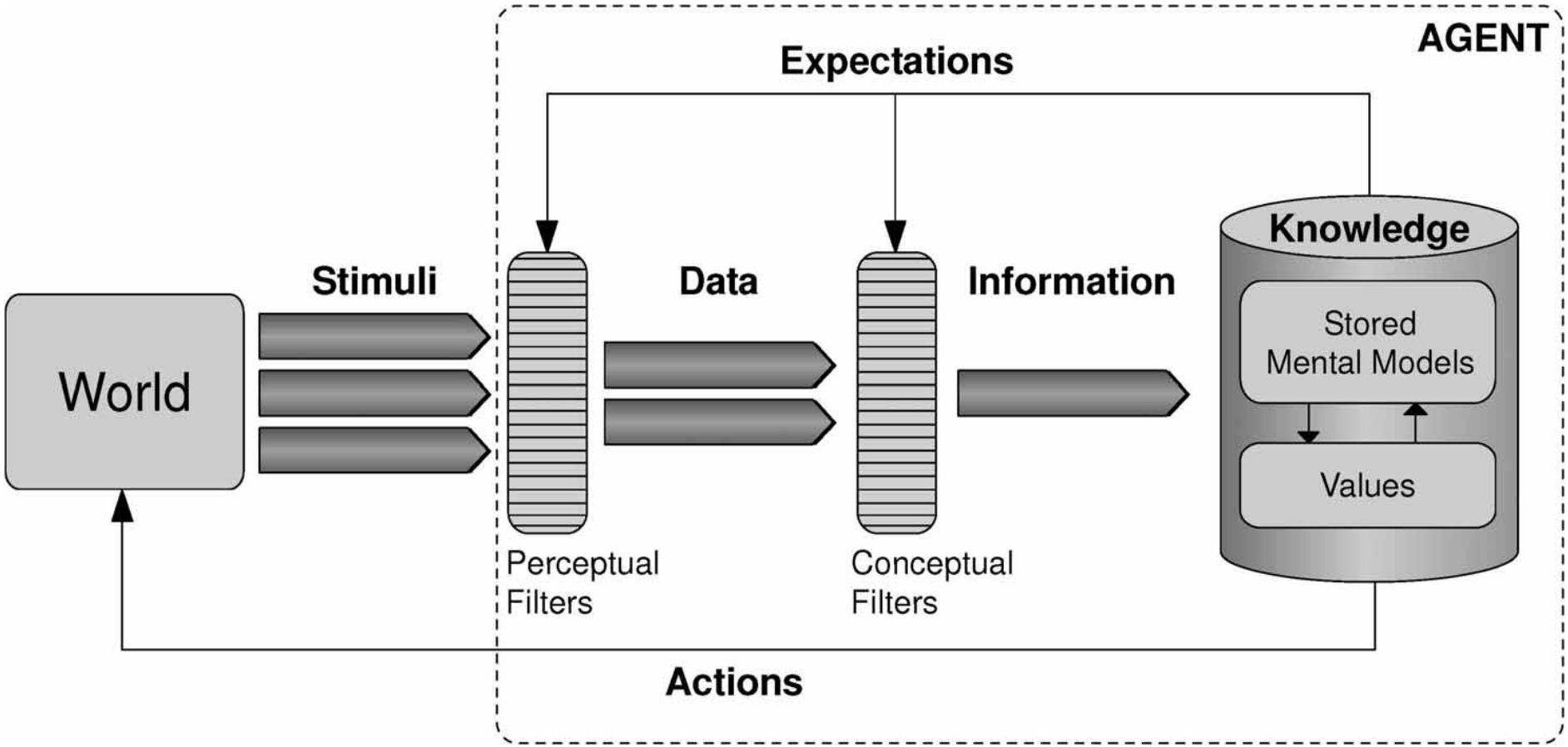
Lane, N. & Martin, W. (2010) The energetics of genome complexity.
Nature, 467, 7318, 929-934.

Slide 2-19: Human Information Processing Model

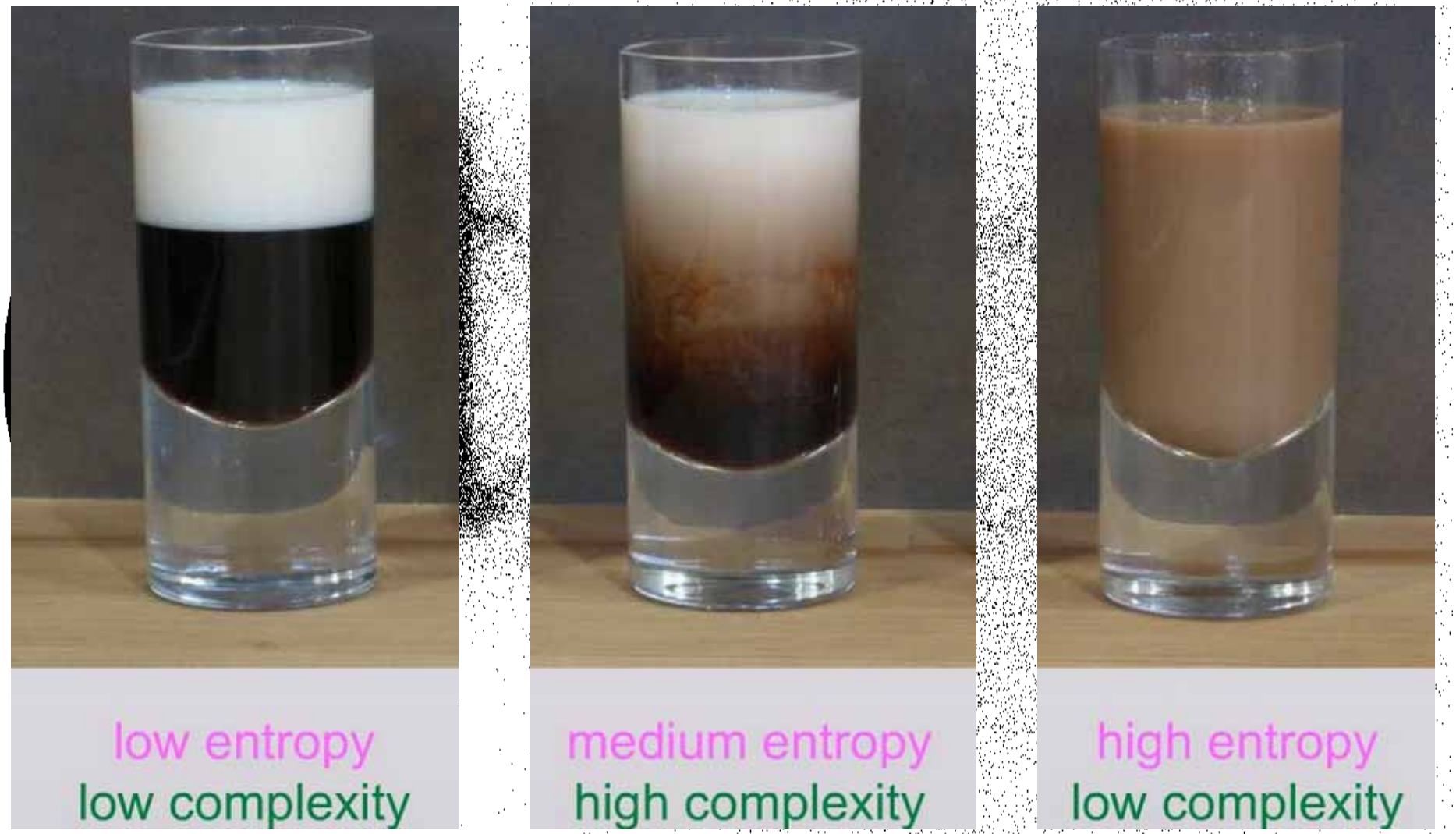


Wickens, C. D. (1984) *Engineering psychology and human performance. Columbus: Merrill.*

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.*



http://www.scottaaronson.com

Σλιδε 2–21: 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."

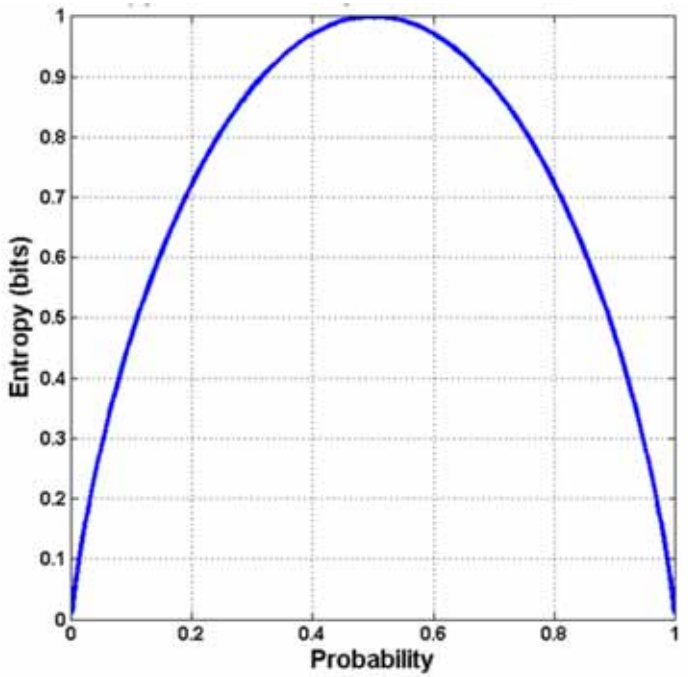
Tribus, M. & McIrvine, E. C. (1971) Energy and Information. *Scientific American, 225, 3, 179-184.*

Slide 2-22: Entropy H as a measure for uncertainty (1/3)

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

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

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



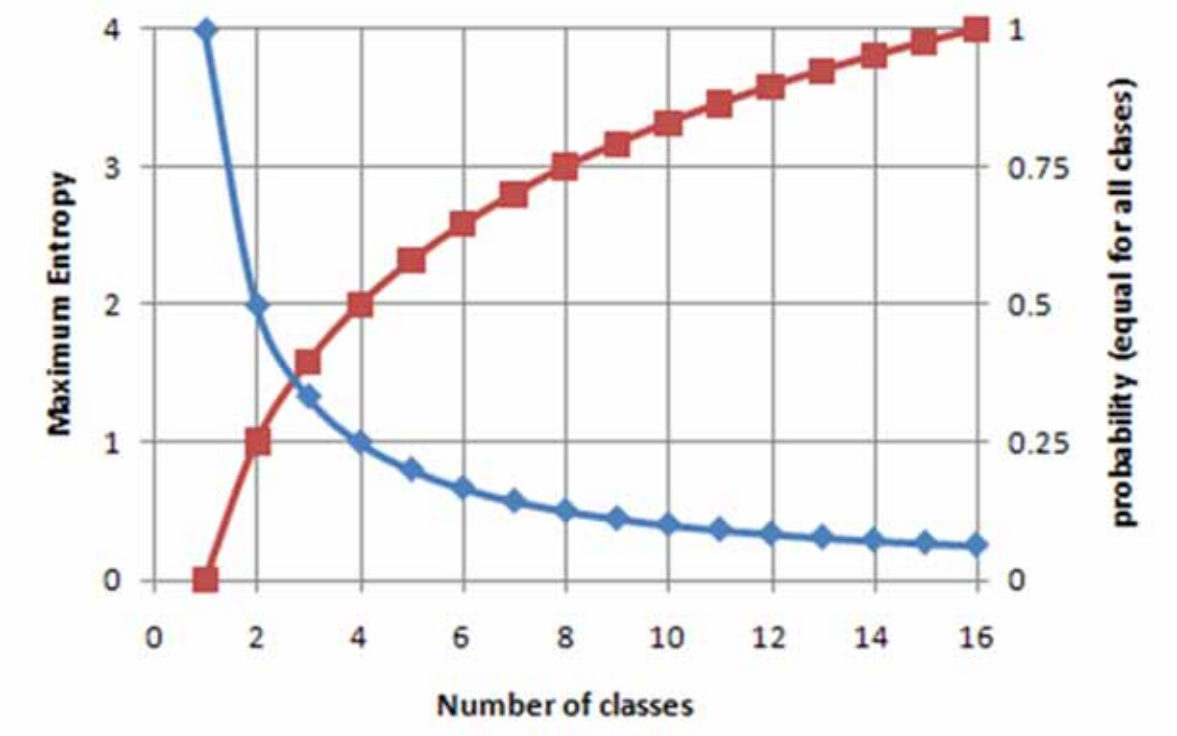
Shannon, C. E. (1948) A Mathematical Theory of Communication. *Bell System Technical Journal, 27, 379-423.*

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

Slide 2-23: A measure for uncertainty (2/3)

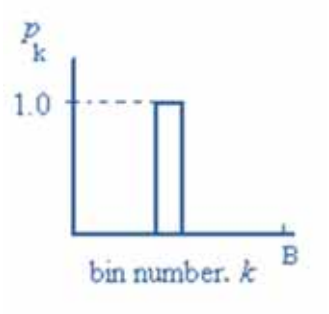
$$\log_2 \frac{1}{p} = -\log_2 p$$

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

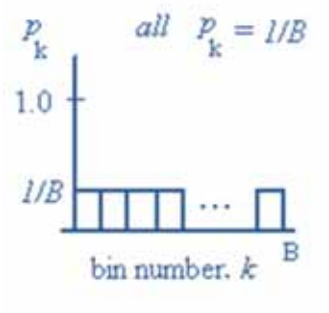


Shannon, C. E. (1948) A Mathematical Theory of Communication. *Bell System Technical Journal, 27, 379-423.*

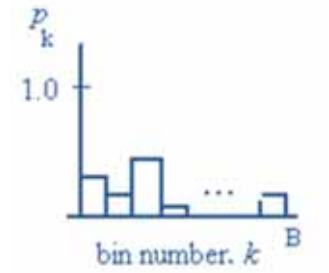
Slide 2-24: Entropy H as a measure for uncertainty (3/3)



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



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



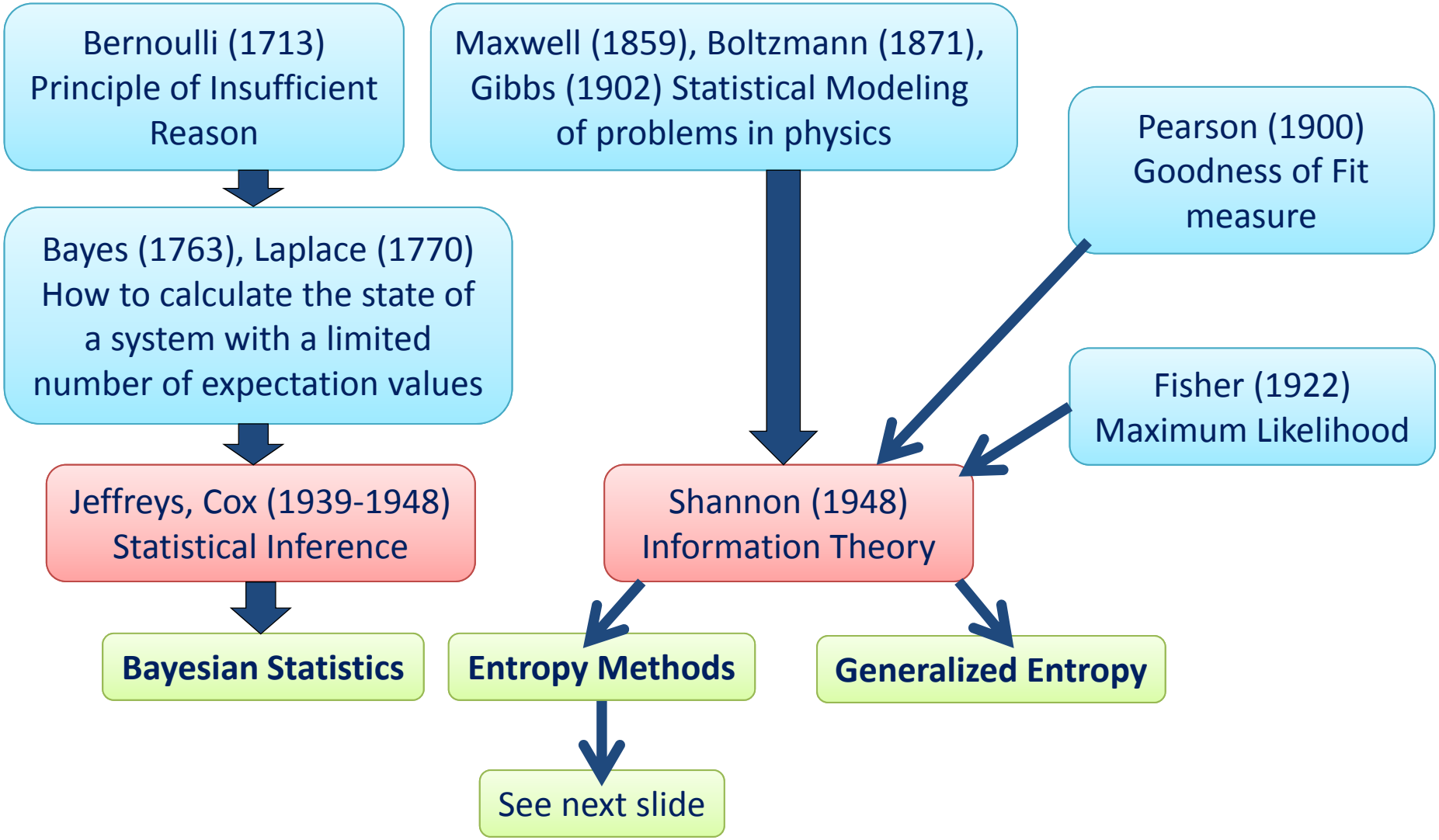
$$H = H_{min} = 0 \qquad H = H_{max} = \log_2 N$$

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, and for modeling etc.;

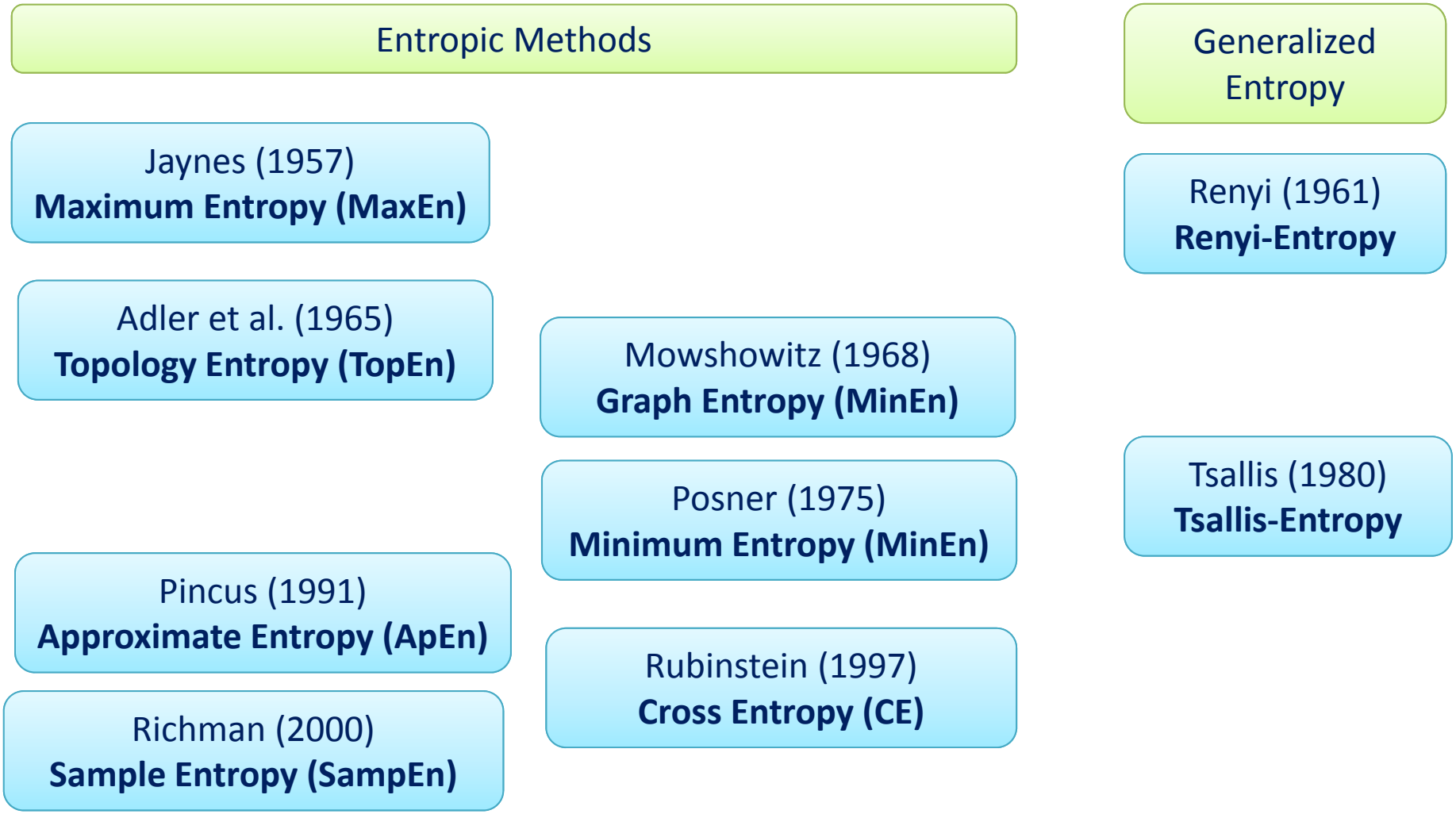
Mayer, C., Bachler, M., Hortenhuber, M., Stocker, C., Holzinger, A. & Wassertheurer, S. 2014. Selection of entropy-measure parameters for knowledge discovery in heart rate variability data. BMC Bioinformatics, 15, (Suppl 6), S2.

Slide 2-25: An overview on the History of Entropy



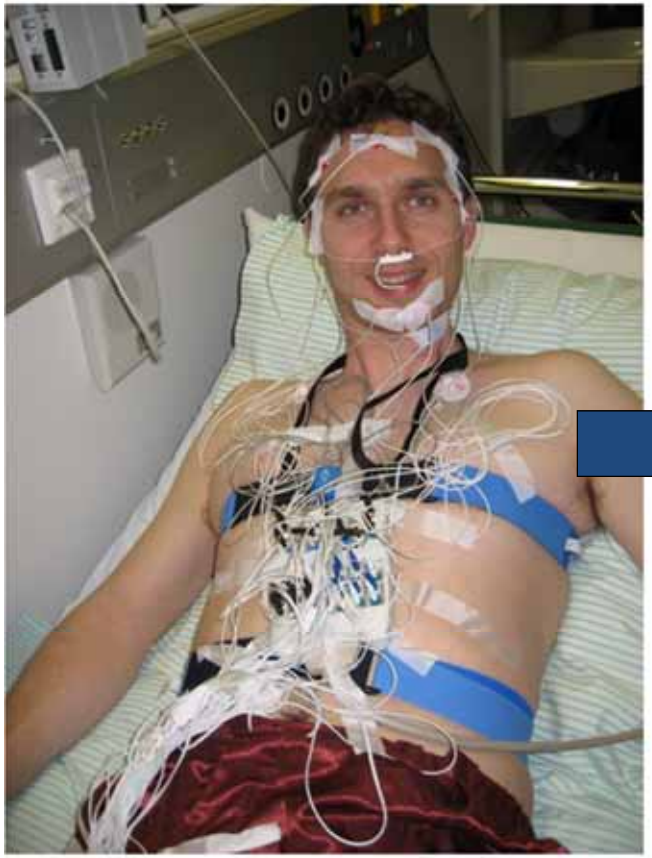
confer also with: Golan, A. (2008) Information and Entropy Econometric: A Review and Synthesis. *Foundations and Trends in Econometrics, 2, 1-2, 1-145*.

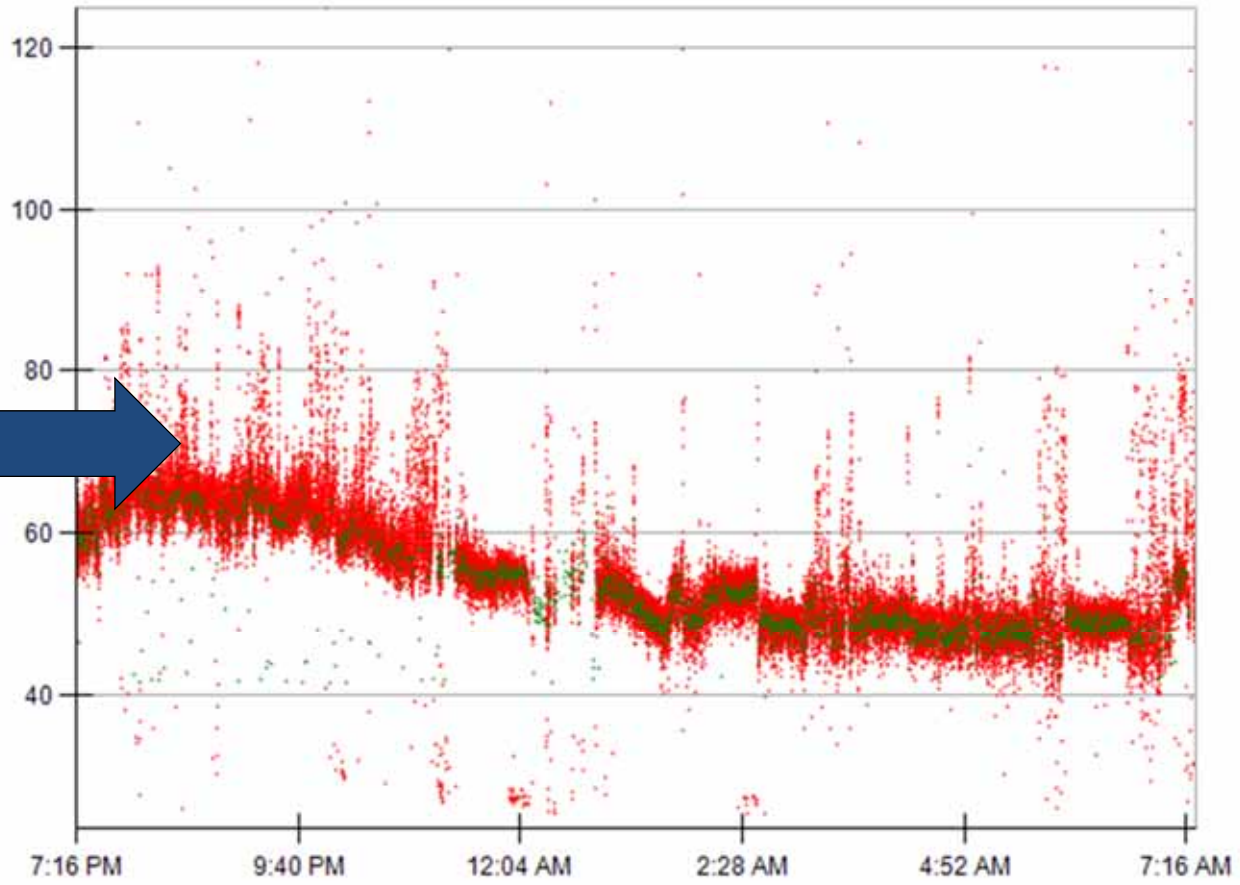
Slide 2-26: Towards a Taxonomy of Entropic Methods



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.

Slide 2-27: 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)

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

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

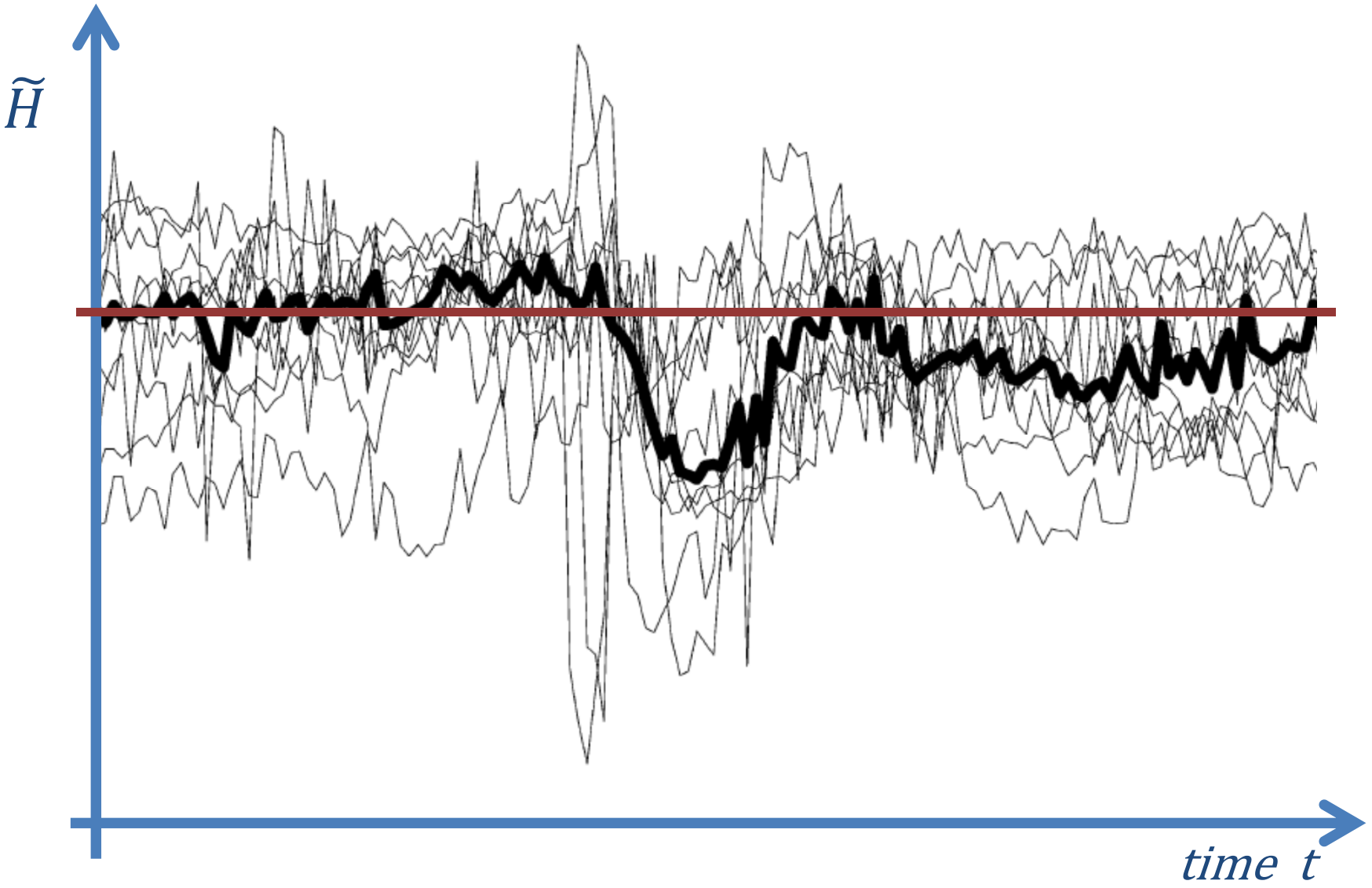
$$\|\vec{X}_i, \vec{X}_j\| = \max_{k=1,2,\ldots,m} \left(\left| x_{(i+k-1)} - x_{(j+k-1)} \right| \right)$$

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

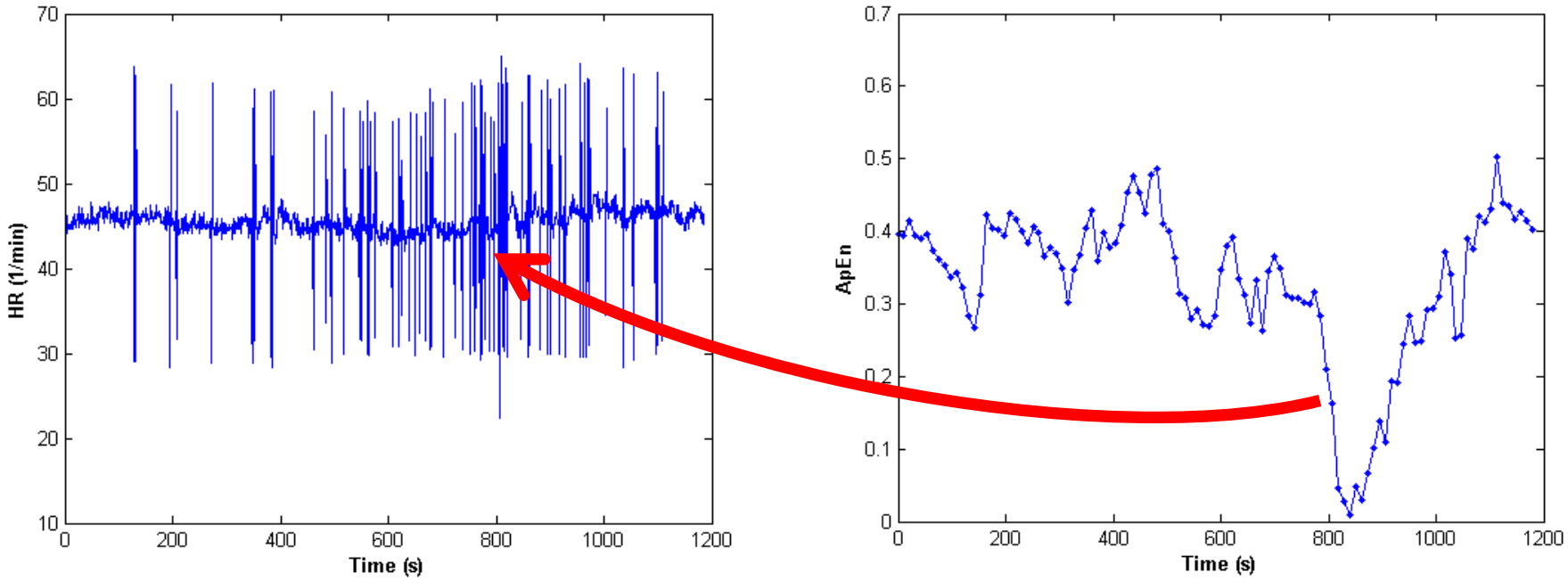
$$C_r^m(i) = \frac{N^m(i)}{N-m+1} \qquad \phi^m(r) = \frac{1}{N-m+1} \sum_{t=1}^{N-m+1} \ln C_r^m(i)$$

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)

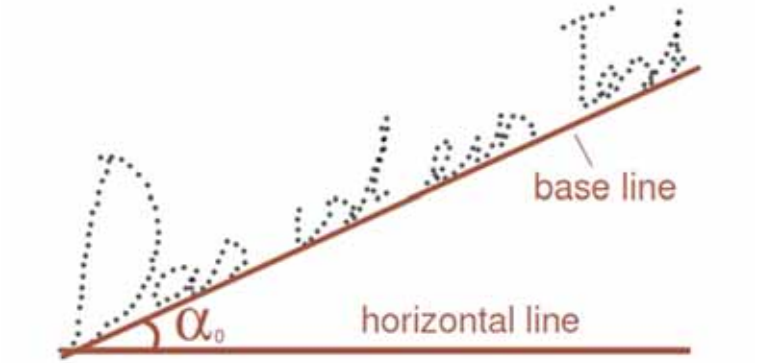


Slide 2-29: ApEn



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.

Example: Skew and Slant correction in Handwriting



$$p_{y,j}(\alpha) = \frac{1}{m+1} \sum_{t=0}^{m} \chi_j(y_\alpha(t))$$

Holzinger, A., Stocker, C., Peischl, B. & Simonic, K.-M. 2012. On Using Entropy for Enhancing Handwriting Preprocessing. Entropy, 14, (11), 2324-2350.

Calculation of Entropy

Algorithm 2 Calculate the entropy $H_{y,\alpha}(X)$ for the projection profiles $p_{y,j}(\alpha)$ for a range of angles α

Require: K = number of data points in X

Require: $\forall k \in \mathbb{N}, 1 \leq k \leq K$: $y_{\min} \leq X[k].y$ **AND** $y_{\max} \geq X[k].y$

Require: $l, \alpha_{\min}, \alpha_{\max} \in \mathbb{N}$ **AND** $-35 \leq \alpha_{\min} < \alpha_{\max} \leq 35$

1: **function** CALCULATEHY($X, K, \alpha_{\min}, \alpha_{\max}, y_{\min}, y_{\max}, l$)
2: $range \leftarrow |\alpha_{\max} - \alpha_{\min}|$
3: $H_y :=$ new vector of size $range$ ▷ Denote the index range from $\alpha_{\min}$ to $\alpha_{\max}$
4: **for** $\alpha = \alpha_{\min} \rightarrow \alpha_{\max}$ **do**
5: $X_\alpha \leftarrow$ ROTATEDATAPOINTS(X, K, α)
6: $w \leftarrow |y_{\max} - y_{\min}| \ / \ l$
7: $p_y \leftarrow$ CALCULATECURRENTPY($X_\alpha, K, y_{\min}, y_{\max}, w$)
8: $H_y[\alpha] \leftarrow 0$
9: **for** $j = 1 \rightarrow l$ **do**
10: $H_y[\alpha] \leftarrow H_y[\alpha] + p_y[j] \cdot \log_2 p_y[j]$
11: **end for**
12: $H_y[\alpha] \leftarrow -1 \cdot H_y[\alpha]$
13: **end for**
14: **return** H_y
15: **end function**

Algorithm 1 Rotating the data points in X for α degree

Require: K = number of data points in X

1: **function** ROTATEDATAPOINTS(X, K, α)
2: $X_\alpha :=$ new vector of size K
3: **for** $k = 1 \rightarrow K$ **do**
4: $X_\alpha[k].x = X[k].x \cdot \cos(\alpha) - X[k].y \cdot \sin(\alpha)$
5: $X_\alpha[k].y = X[k].x \cdot \sin(\alpha) + X[k].y \cdot \cos(\alpha)$
6: $X_\alpha[k].p = X[k].p$
7: **end for**
8: **return** X_α
9: **end function**

Holzinger, A., Stocker, C., Peischl, B. & Simonic, K.-M. 2012. On Using Entropy for Enhancing Handwriting Preprocessing. Entropy, 14, (11), 2324-2350.

$\widetilde{H}$ …

- … is **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!

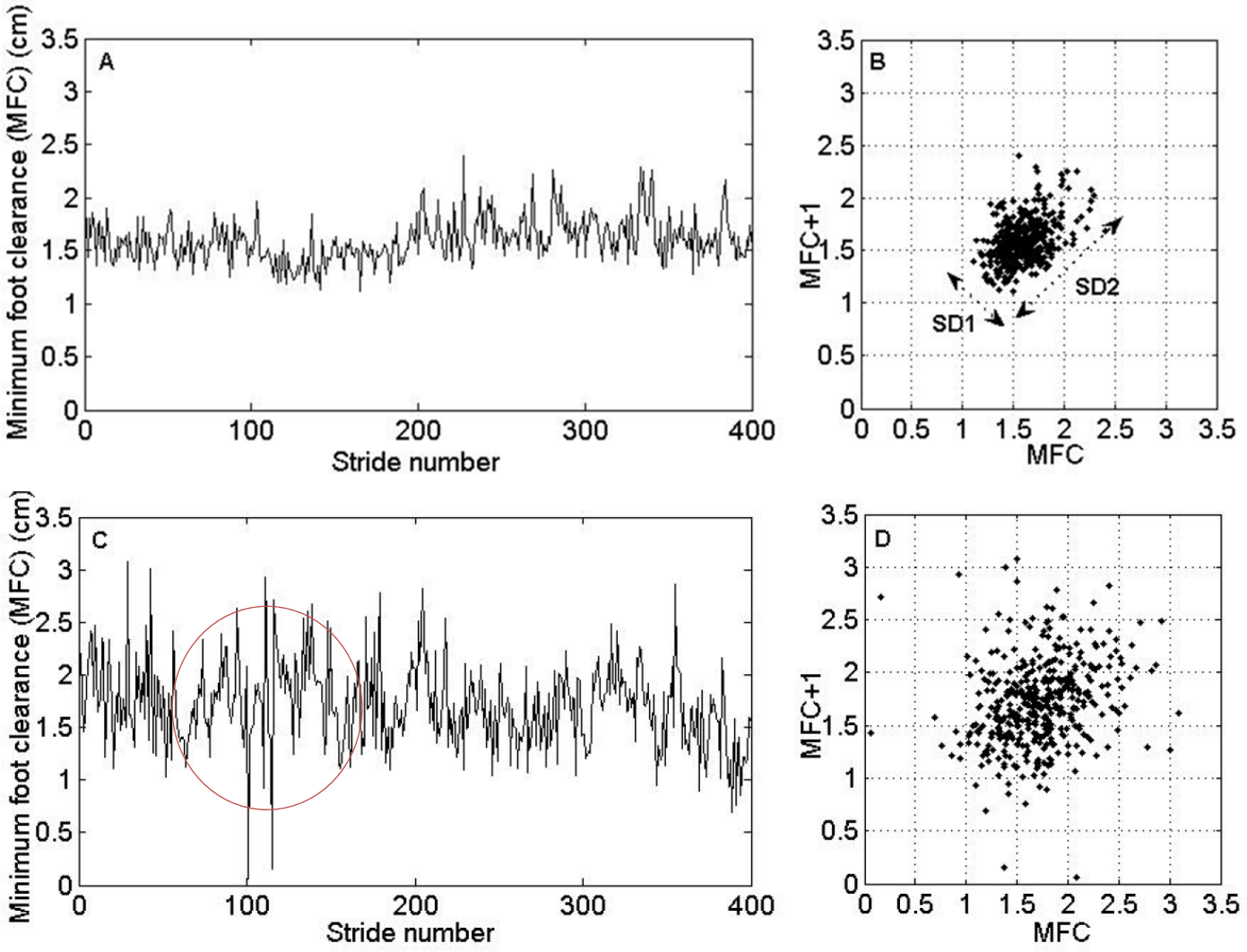
Sample Questions (1)

- 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 is the difference between the perceptual space and the computational space in terms of data, information and knowledge?
- 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?

Back-up Slide: Poincare Plot for gait analysis

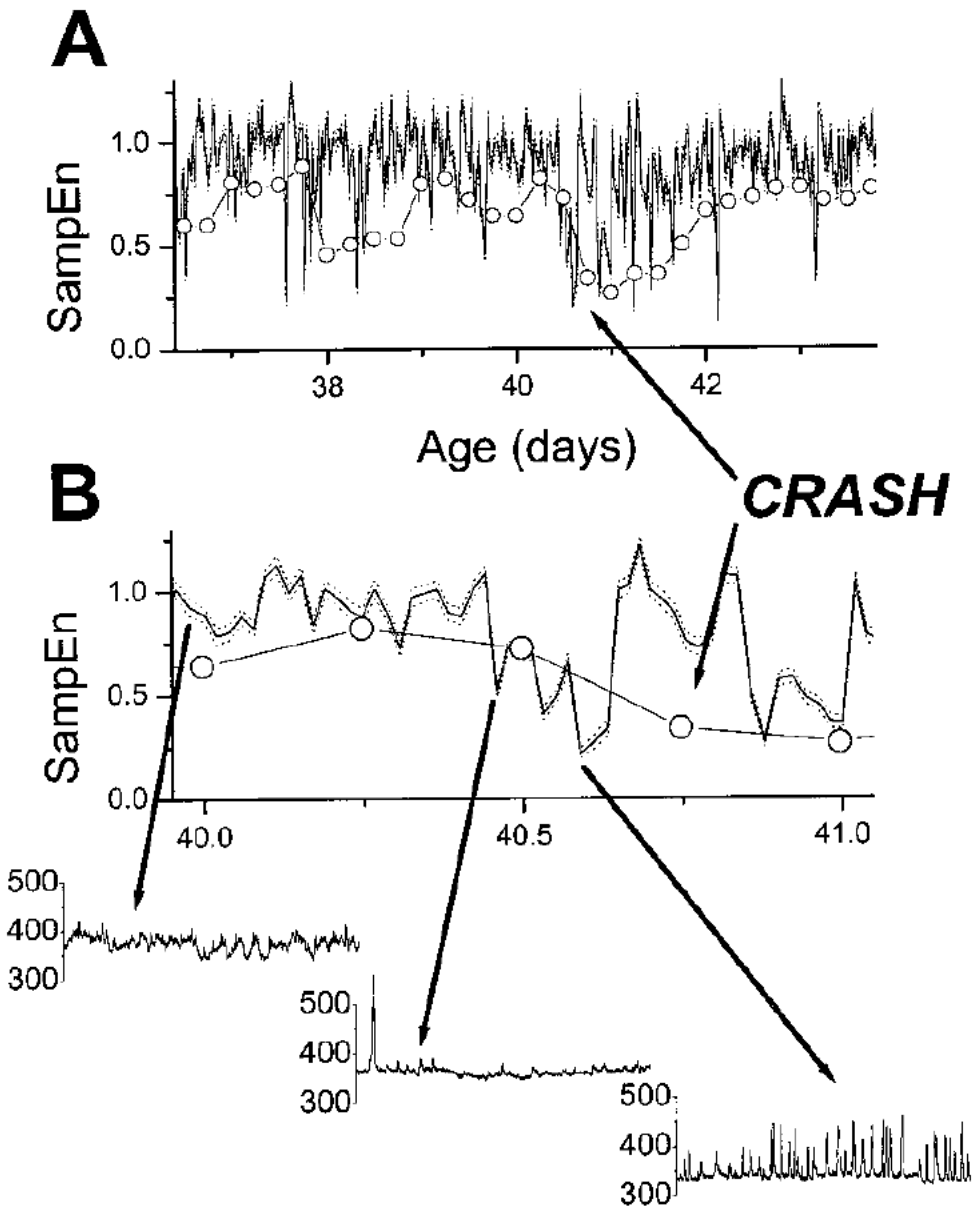


Khandoker, A., Palaniswami, M. & Begg, R. (2008) A comparative study on approximate entropy measure and poincare plot indexes of minimum foot clearance variability in the elderly during walking. *Journal of NeuroEngineering and Rehabilitation, 5, 1, 4.*

Sample Exam Questions – Yes/No Answers

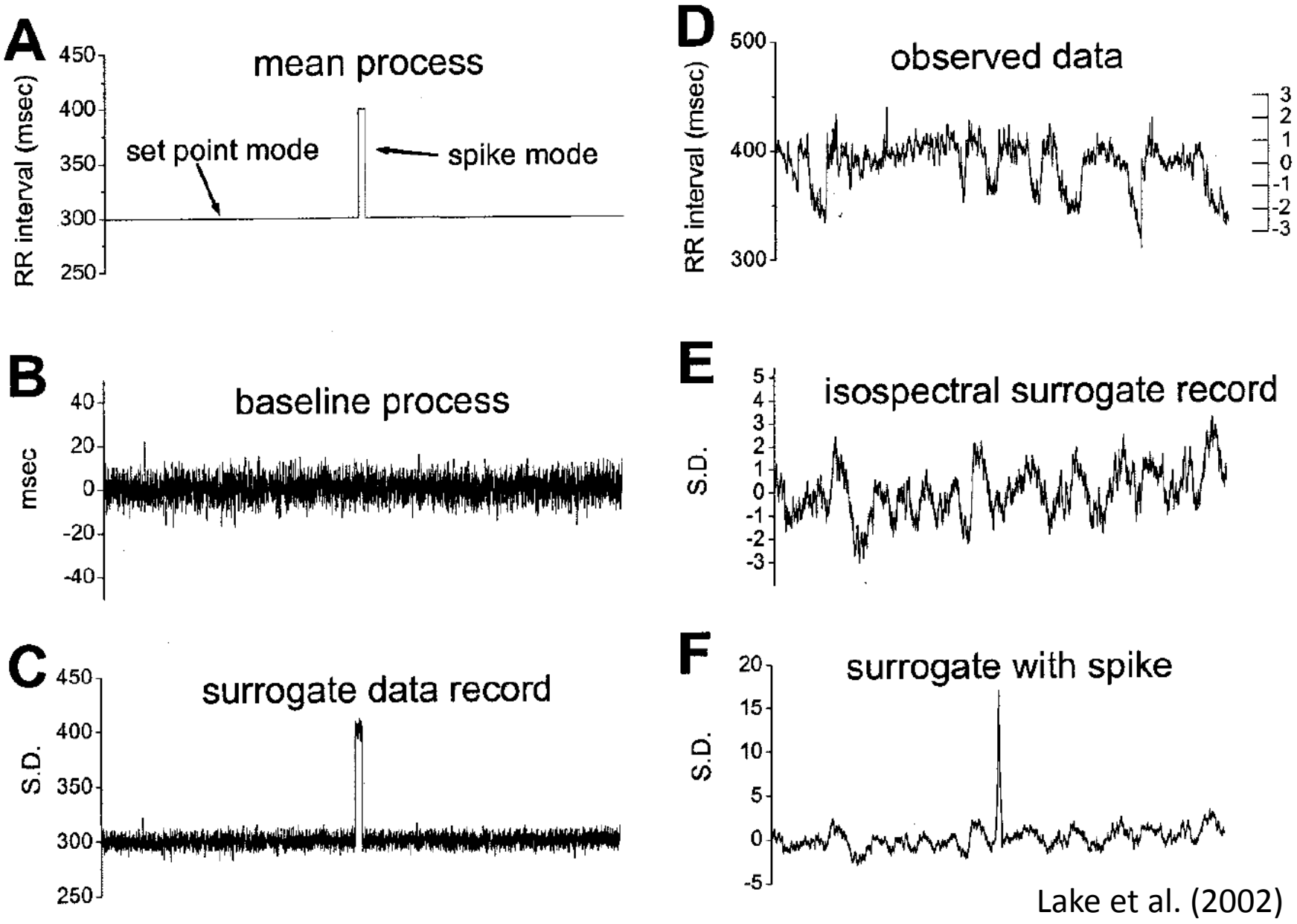
01	An array is a composite data type on physical level.	❒ Yes ❒ No	2 total
02	In a Von-Neumann machine "List" is a widely used data structure for applications which do not need random access.	❒ Yes ❒ No	2 total
03	The edges in a graph can be multidimensional objects, e.g. vectors containing the results 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)			

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:

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

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

$$C_r^m(i)=V^m(i)/(N-m+1) \quad (3)$$

$$where \quad V^m(i)=no.of\ d[X(i),X(j)]\le r$$

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

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

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)

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

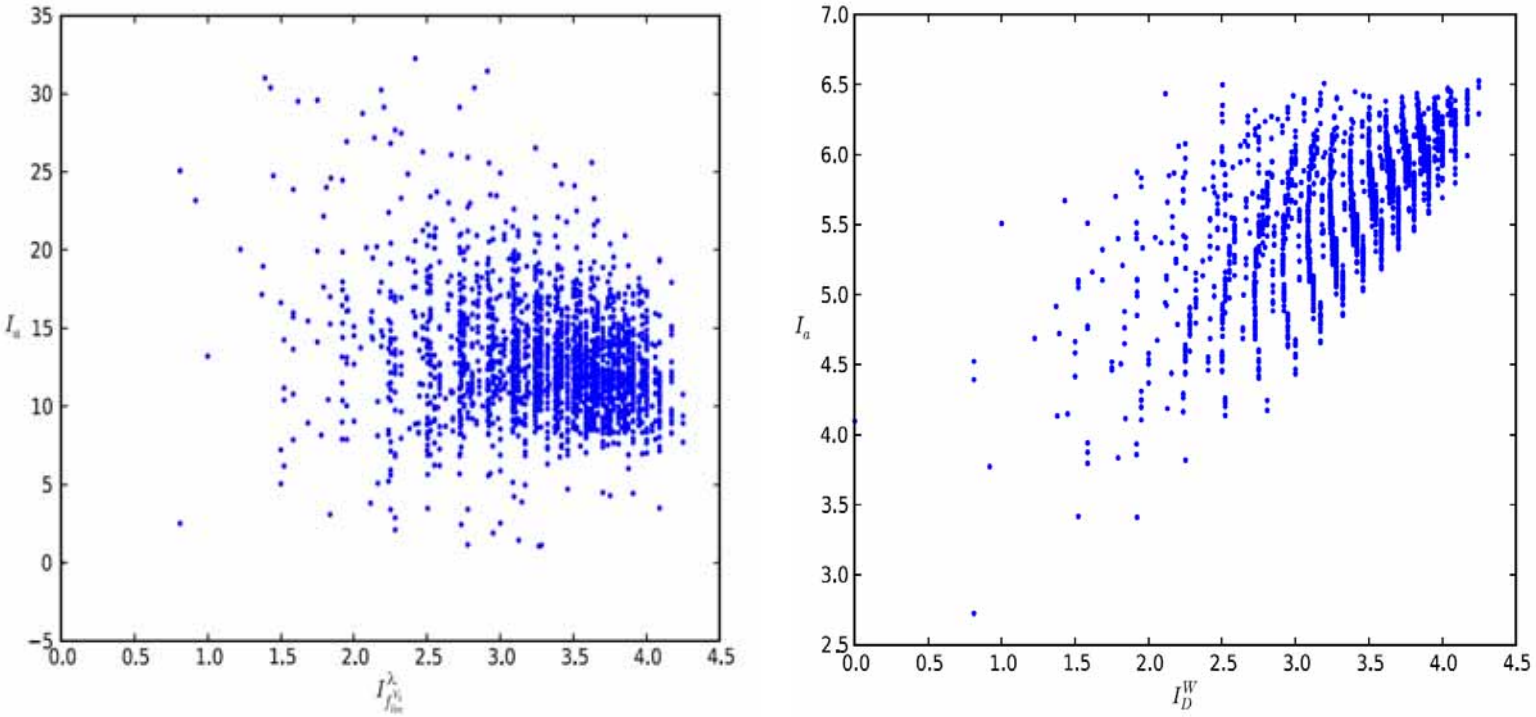
6) SampEn value for a finite data length of *N* can be estimated:

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

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

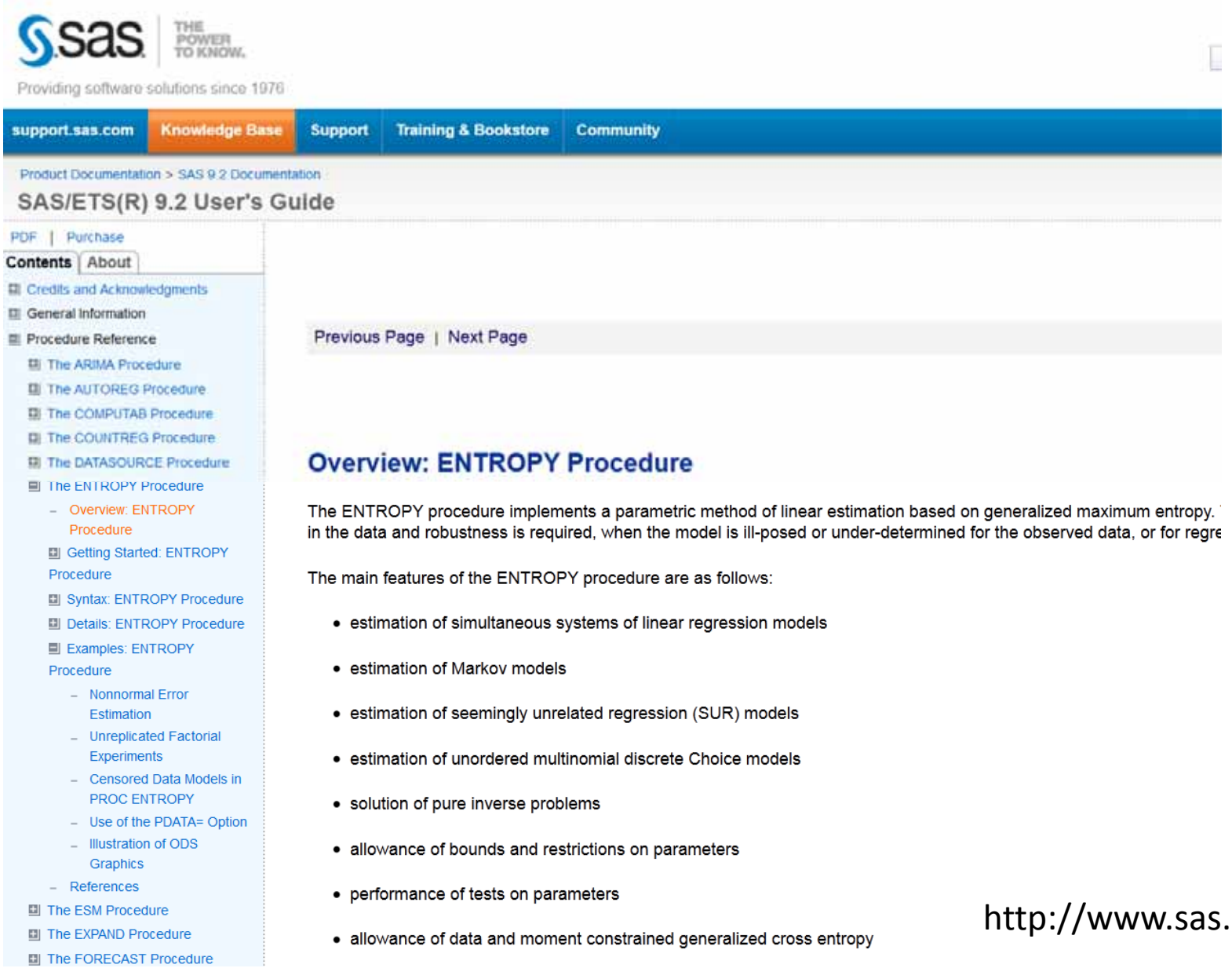
http://www.statistik.at

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)



http://www.sas.com

Backup Slide: Example Tool for large data sets - Hadoop

Top | Common | HDFS | MapReduce

- About
  - Welcome
  - Mailing Lists
  - Who We Are?
  - Who Uses Hadoop?
  - Buy Stuff
  - Sponsor Apache
  - Sponsors of Apache
  - Privacy Policy
  - Bylaws
- Sub-Projects
- Related Projects

built with Apache Forrest

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 compu machines, each offering local computation and storage. Rather than rely on hardware to deliver high-avaiability, the library itself is desig a cluster of computers, each of which may be prone to failures.

The project includes these subprojects:

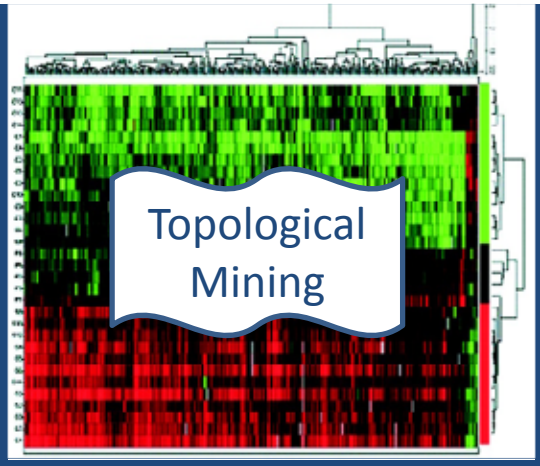
- **Hadoop Common**: The common utilities that support the other Hadoop subprojects.
- **Hadoop Distributed File System (HDFS™)**: A distributed file system that provides high-throughput access to application data.
- **Hadoop MapReduce**: A software framework for distributed processing of large data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

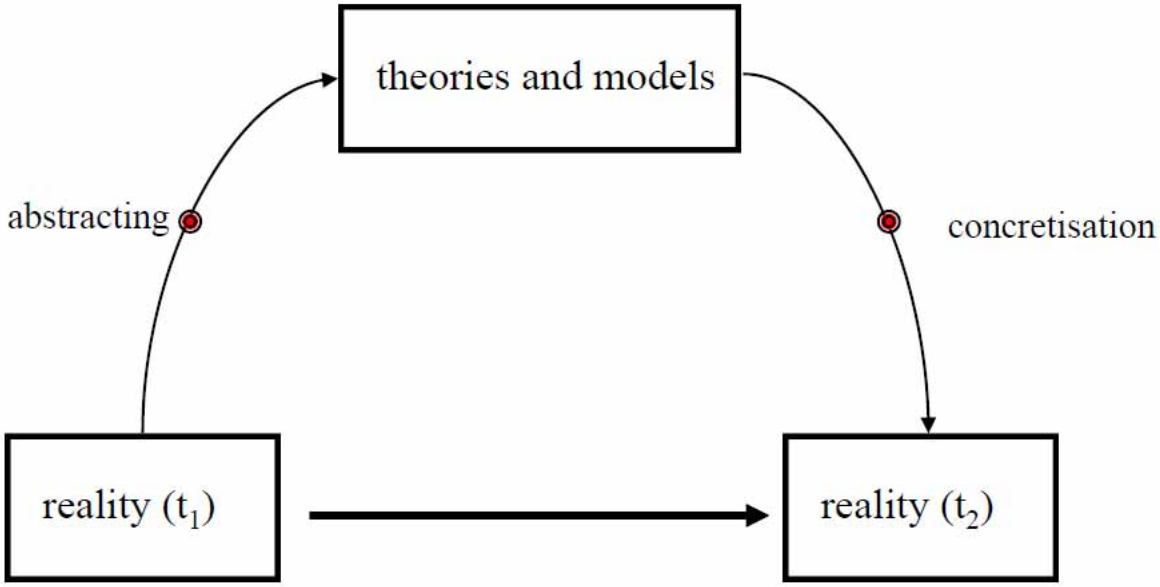
Backup Slide: Methods for Mining ..



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)

Backup Slide: Excursion: How to get theories?



positivism :

{theory, model} $\notin$ reality

reality (t_1) $\approx$ reality (t_2)

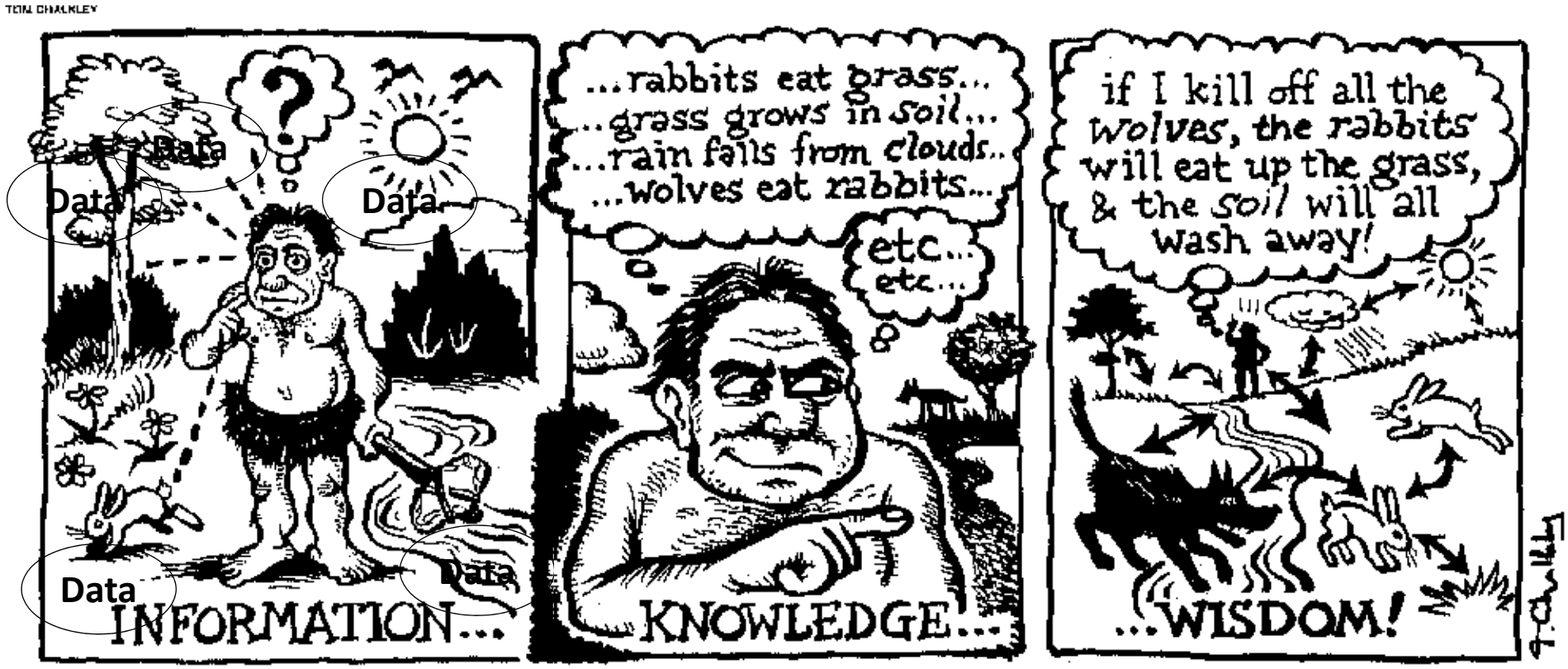
constructionism :

{theory, model} $\in$ reality

reality (t_1) $\neq$ reality (t_2)

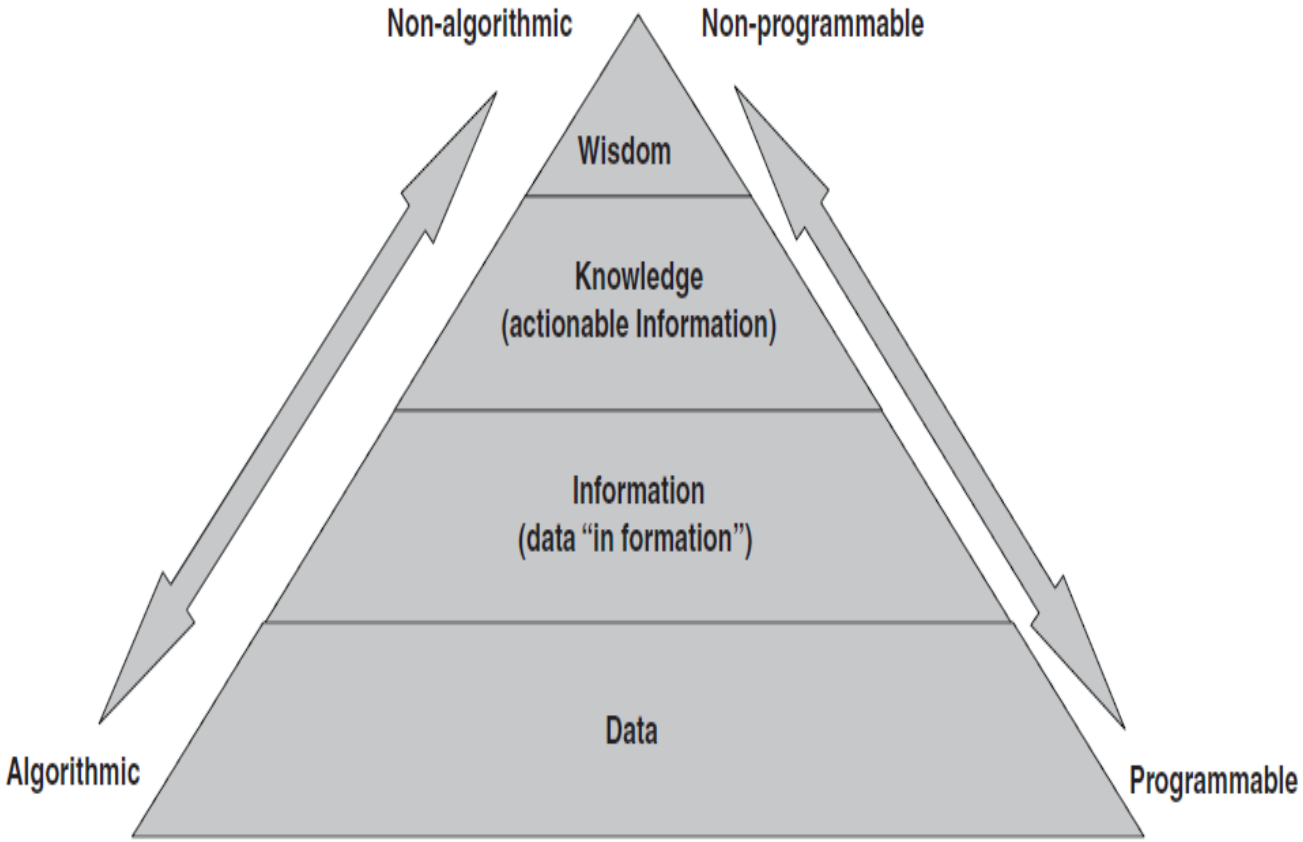
Rauterberg, M. (2006) HCI as an engineering discipline: to be or not to be.
African Journal of Information and Communication Technology, 2, 4, 163-184.

Backup Slide: The DIKW Model (1/4)



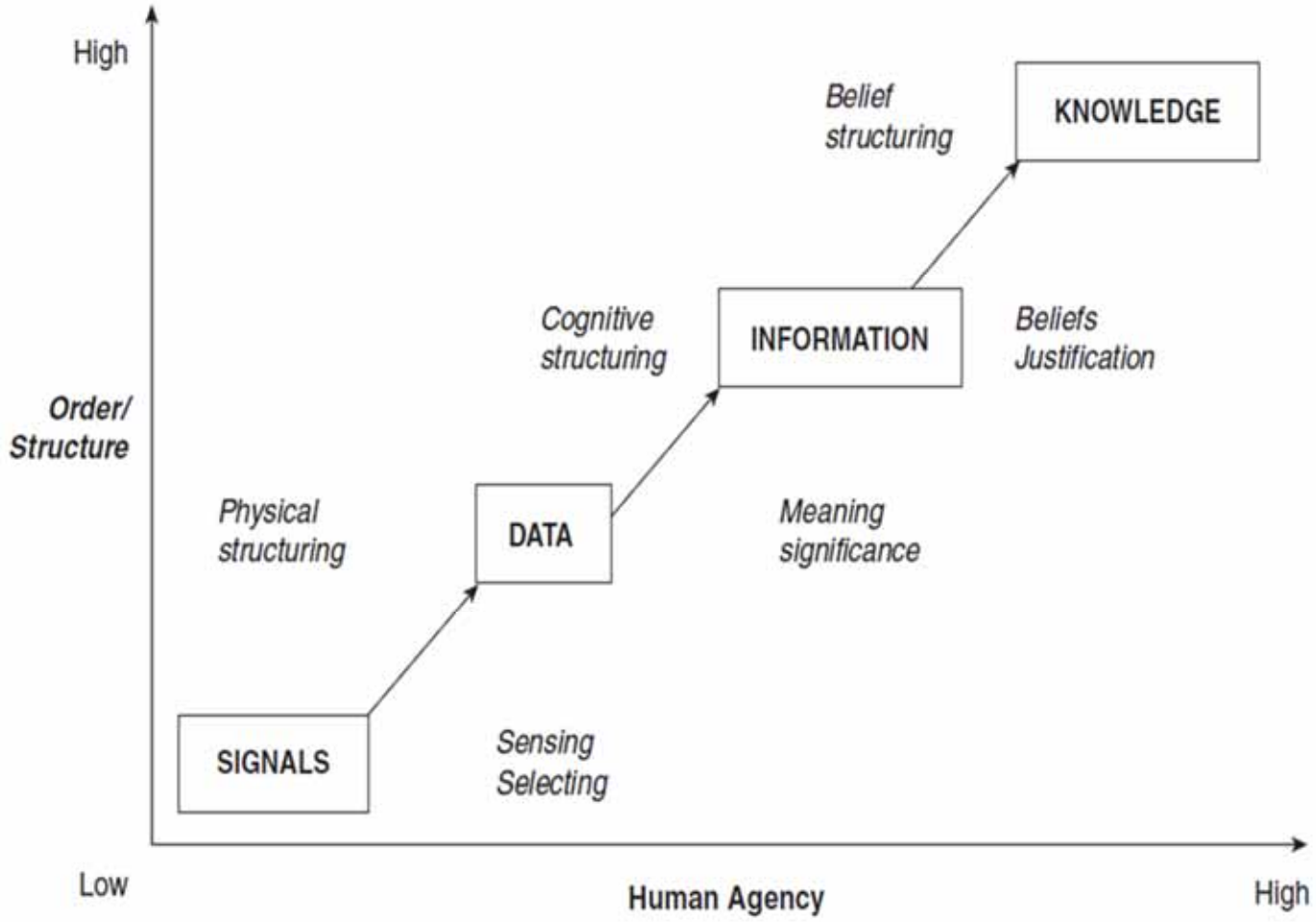
Cleveland H. "Information as Resource", The Futurist, December 1982 p 34-39.

Backup Slide: The DIKW Model (2/4)



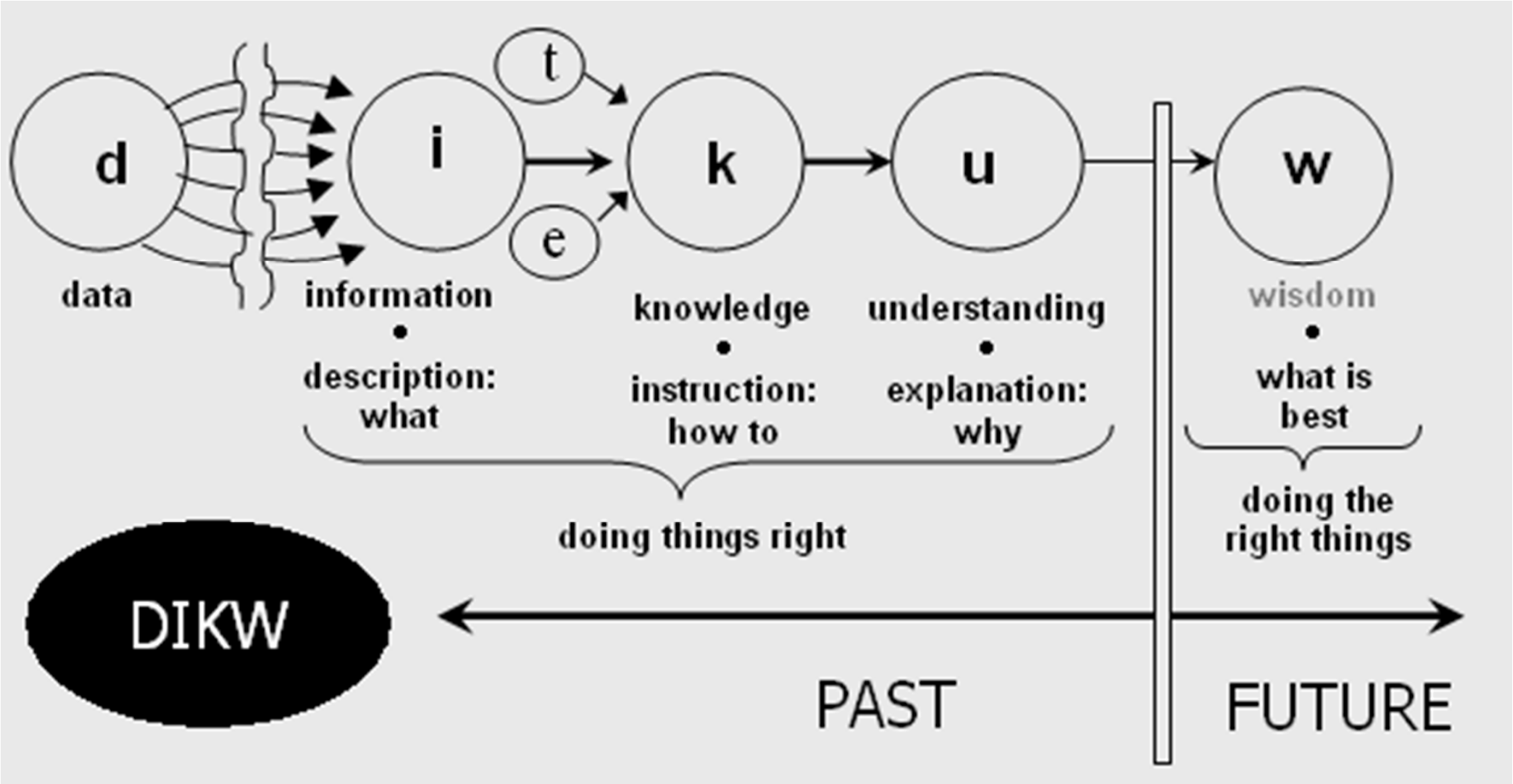
Rowley, J. (2007) The wisdom hierarchy: representations of the DIKW hierarchy. *Journal of Information Science, 33, 2, 163-180.*

Backup Slide: The DIKW Model (3/4)



Rowley, J. (2007) The wisdom hierarchy: representations of the DIKW hierarchy. *Journal of Information Science, 33, 2, 163-180.*

Backup Slide: The DIKW Model (4/4)



Source: Public Domain http://en.wikipedia.org/wiki/DIKW

For critic on this model see for example: Fricke, M. (2009) The knowledge pyramid: a critique of the DIKW hierarchy. Journal of Information Science, 35, 2, 131-142.